UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DISEÑO Y DESARROLLO DE UN SISTEMA PARA LA ASOCIACIÓN AUTOMÁTICA DE DELITOS BASADOS EN MODELOS SIMILITUD SEMÁNTICA TEXTUAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SEBASTIÁN IGNACIO SANTANA RUIZ

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
LUIS ABURTO LAFOURCADE
CAROLINA SEGOVIA RIQUELME

SANTIAGO DE CHILE
2018

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: SEBASTIÁN IGNACIO SANTANA RUIZ
FECHA: 2018
PROF. GUÍA: RICHARD WEBER HAAS

DISEÑO Y DESARROLLO DE UN SISTEMA PARA LA ASOCIACIÓN AUTOMÁTICA DE DELITOS BASADOS EN MODELOS SIMILITUD SEMÁNTICA TEXTUAL

La Fiscalía de Chile o Ministerio Público es una institución autónoma, cuya función es dirigir, a través de sus fiscales y en forma exclusiva, la investigación de los hechos que pueden ser constitutivos de delitos. El problema es que sólo el año 2017 el Ministerio Público recibió más de 1.3 millones de denuncias, las cuales deben deben ser investigadas con un acotado cuerpo de fiscales y analistas, lo que se traduce en que para dar respuesta a esta cifra y bajo el supuesto de una homogeneidad en la carga de trabajo, cada fiscal debería atender, investigar y si es posible llevar a juicio 7 causas diarias.

En el estado del arte existen trabajos que vinculas delitos a través de modelos de clustering y clasificación obteniendo buenos resultados, no obstante esto supone la existencia de bases de datos con campos definidos y bien pobladas, lo cuál difiere del caso en Chile dada la forma en la que se recogen las denuncias. En un intento por contribuir al problema del Ministerio Público, es que se ha desarrollado un modelo para la vinculación automática de delitos basado en métricas de similitud semántica textual derivada de modelos de aprendizaje de máquina. Para esto, se ha diseñado un proceso que comienza por la recuperación de documentos a través de queries mediante Latent Semantic Indexing (LSI), para luego computar y analizar la asociación de causas recuperadas a través de modelos de similitud semántica textual, en este caso Doc2Vec. Finalmente, y en el caso de que el resultado brinde asociaciones de causas muy numerosas, se propone su descomposición a través de modelos de tópicos, en este caso y por simplicidad, Latent Dirichlet Allocation (LDA).

En primer lugar, en un conjunto de 3.803 causas se realizó el ejercicio de comparar agrupaciones que establecidas por nuestro sistema con causas que el personal del Ministerio Público se encontraba investigando. El resultado, es que a partir de una consulta se encontraron 7 agrupaciones, que sumaban 66 causas en total y dentro de las cuales se encontraban 4 de los 16 delitos que el Ministerio Público investigaba. Luego, en el mismo conjunto de causas se analizó que causas pudiesen estar relacionadas con las que se investigaban. En 56 causas analizadas, 9 fueron validadas como delitos con un modo de comisión similar a los que se investigaban, lo que en términos de Precision corresponde a un 19 %.

Se ha desarrollado una metodología que ha demostrado funcionar tanto para la agrupación de denuncias a partir de términos de búsqueda cómo para la asociación de nuevas causas a delitos en investigación, en donde la elección de los modelos ha resultado ser efectiva, contribuyendo al análisis de un gran volumen de denuncias de forma automática Los resultados son prometedores dada la complejidad del problema y se proponen nuevos desarrollos para complementar esa incipiente versión del sistema para asociación de delitos, donde además cabe destacar que no existen registros en la literatura de trabajos de vinculación criminal basado exclusivamente en datos no estructurados.

A quién fue el primero en alentarme en esta dirección

... gracias Gustavo.

Agradecimientos

En primer lugar a mi madre Evelyn, quién me ha acompañado, cuidado y amado incondicionalmente. A mi padre Juan, por su cariño, preocupación y guía. Gracias por ayudarme a desarrollar una visión esperanzadora de la vida.

A mi abuelo Juan, por ser un segundo padre y darme un infancia feliz.

Al profesor Richard, por la oportunidad de incluirme en esta aventura, la confianza, el apoyo y la preocupación.

A los miembros de esta comisión: Luis, Carolina y Patricio. Gracias por guiarme durante este proceso y darme de su tiempo cada vez que lo necesite.

Finalmente, te agradezco a tí, ya que vuelves a dar vida a este trabajo a través de su lectura, ¡Gracias!

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Containing the spread of crime in urban societies remains a major challenge. Empirical evidence suggests that, if left unchecked, crimes may be recurrent and proliferate. On the other hand, eradicating a culture of crime may be difficult, especially under extreme social circumstances that impair the creation of a shared sense of social responsibility.

Maria D'Orsogna y Matjaž Perc, 2014

1.1. Antecedentes generales

La Fiscalía de Chile o Ministerio Público es una institución autónoma, cuya función es dirigir en forma exclusiva la investigación de las causas criminales, para lo cual debe dar las órdenes que corresponda a las policías y a otros servicios para que desarrollen las acciones investigativas, tanto para esclarecer los hechos denunciados, como para acreditar la participación o inocencia del imputado; ejercer, cuando resulte procedente, la acción penal pública, formulando acusación e instando por resolver adecuada y oportunamente los diversos casos penales y adoptar las medidas necesarias para la atención y protección de víctimas y testigos. La Fiscalía puede iniciar la investigación de un hecho que reviste caracteres de delito por denuncia; por querella; o de oficio, es decir, por propia iniciativa, en este último caso, cuando los fiscales toman conocimiento personal o presencial de la comisión de un delito. Las denuncias se pueden ejercer ante cualquier Comisaría de Carabineros o ante cualquier cuartel de la PDI. También se pueden presentar las denuncia directamente en la fiscalía local del lugar donde ocurrieron los hechos, o en aquella que el denunciante estime conveniente según el lugar donde se encuentre, finalmente, las denuncias también puede ser presentadas ante un tribunal con competencia criminal.

1.1.1. Caracterización del problema

Según datos del Centro de Estudios y Análisis del Delito, en la década del año 2007 al año 2017 los casos policiales por Delitos de Mayor Connotación Social (DMCS)[1] sufrieron una baja del 7,19 % en la tasa cada 100.000 habitantes, pasando de 3.324 casos policiales cada 100.000 habitantes en el año 2007 a 3.085 el 2017. El problema es que entre el año 2008 y el 2014 las causas ingresadas al Ministerio Público fueron constituidas en un 47 % por delitos donde el autor del delito es desconocido (figura conocida como Imputado Desconocido), donde las salidas judiciales de causas con imputado desconocido llega sólo al 13.4 % vs un 69 % de las causas donde el imputado es conocido[2].

Por otro lado, sólo en el año 2017 el Ministerio Público recepcionó 1.323.324 denuncias de delitos. Para atender ese volumen de denuncias, la Fiscalía cuenta con 751 fiscales a nivel nacional, lo que implica que ante una homogeneidad en la carga de trabajo, cada fiscal debería atender, investigar y si es posible llevar a juicio 1.762 causas anuales o 7 causas diarias[3]. Este desproporcionado volumen de causas para un acotado cuerpo de investigación provoca que dar respuesta oportuna y eficaz a cada una de las denuncias sea imposible en la práctica. Más aún, si se espera que este organismo de respuesta a la cifra de delitos cuyo autor es recurrente y desconocido, es decir, un delincuente que comete más de un delito, pero sigue impune y libre, es necesario indagar, vincular e investigar en detalle las denuncias recepcionadas que poseen un relato similar[4]. La Fiscalía posee apoyo tecnológico para esta maratónica labor, de hecho se encuentran el pleno proceso de la implementación de una plataforma denominada Sistema Integrado de Monitoreo y Análisis Criminal (SIMAC), el cual es un repositorio de información que permite estructurar los datos relevantes para caracterizar un delito, de manera que sea posible realizar la búsqueda de delitos que son similares entre sí a través de consultas al sistema de información. No obstante, para poblar de información y datos dicho sistema, es necesario que analistas completen los campos de información para cada delito, lo cual nos remite nuevamente al problema de la gran cantidad de registros que llegan al Ministerio Público.

1.1.2. Justificación de este trabajo

En otro orden de cosas, en los últimos 5 años las técnica, métodos y modelos de Text Mining han mostrado grandes resultados en la extracción de conocimiento desde textos. En particular, el desarrollo de los Word Embeddings ha permitido generar una representación

[1] Acorde a lo señalado en la página de Carabineros de Chile, estos corresponden a los delitos denominados como "Delitos Violentos"(Robo con Violencia, Robo con Intimidación, Robo por Sorpresa, Lesiones, Homicidio y Violación), y "Delitos Contra la Propiedad"(Robo de Vehículo Motorizado, Robo de Accesorios de Vehículos, Robo en Lugar Habitado, Robo en Lugar no Habitado, Otros Robos con Fuerza y Hurto). Recuperado en Agosto del 2018.

[2] Cifras del documento elaborado por la Fundación Paz Ciudadana en su presentación Radiografía del sistema de seguridad y justicia, Agosto 2015.

[3] En el año 2017 hubo 249 días laborales.

[4] El año 2016 el Ministerio Público implementó el denominado Sistema de Análisis Criminal y Focos Investigativos (SACFI), el cual busca agrupar o reunir un número importante de causas de la misma naturaleza o figura delictiva[5] e indagarlas bajo un solo proceso, cuya figura tiene el nombre de Foco de Investigación.

vectorial de los documentos que permite capturar la semántica y el contenido que abarcan los documentos escritos. Es por esto que surge la idea de intentar contribuir al problema que enfrenta el Ministerio Público desde una perspectiva del procesamiento automático de las denuncias con el objetivo de aumentar la capacidad de análisis e investigación de los analistas y fiscales del Ministerio Público.

En la actualidad, los focos de investigación del SACFI se constituyen en base a información exógena al sistema, es decir, por mandatos directos, información de fuentes externas o priorización en base a un aumento considerable en un determinado delito. Luego de eso se realiza el seguimiento del foco, el cuál consiste en la validación y agregación de causas que puedan estar vinculadas a los delitos ya relevados. El seguimiento del foco de investigación se hace a través del estudio de otras causas, muchas veces sin indicios o un sistema de priorización, lo que implica que analistas y fiscales no poseen una herramiento que les indique si existen causas que posean un alto nivel de similitud con las causas que se investigan. Así mismo, no existe una herramienta que analice de forma automática los delitos y en base a la similitud del método de comisión del delito (o *modus operandi*) determine un potencial foco de investigación.

Así, en el presente trabajo se plantea un intento por asociar delitos a través de la similitud semántica textual, a partir de las denunicas en poder del Ministerio Público con el objetivo de contribuir a las labores de investigación que lleva a cabo esta institución.

1.2. Objetivos

1.2.1. Obejtivo general

« Desarrollar un modelo para la vinculación automática de delitos basado en métricas de similitud semántica textual derivada de modelos de aprendizaje de máquina »

1.2.2. Objetivos específicos

Para cumplir con el propósito establecido se han fijado 5 objetivos específicos que definen el alcance y metodología de este trabajo.

1. Definir la figura delictiva en la que se desempeñará el trabajo. Este objetivo es de particular relevancia, ya que no todos los delitos son investigados de igual manera y cada tipo de delito posee sus propios desafíos, fuentes de información y procedimientos para su debida investigación.
2. Implementar algoritmos de recuperación de documentos basado en queries.
3. Implementar y evaluar los resultados de los algoritmos de similitud semántica textual.
4. Desarrollar visualizaciones que permitan entender la información contenida en los con-

juntos de recomendación.

5. Validar los resultados obtenidos con expertos del Ministerio Público.

1.3. Hipótesis del trabajo

A juicio del autor de este trabajo, la principal innovación en el estado del arte y los mecanismos actuales con los que cuenta el Ministerio Público para la búsqueda, análisis e investigación de los delitos es la implementación de los más avanzados algoritmos de similitud semántica textual. Lo anterior permite avanzar en el diseño de búsquedas inteligentes de delitos que pueden estar potencialmente vinculados o que posean un *modus operandi* común. Esto permitiría generar conjuntos de causas que posean un alto nivel de similitud en el relato del hecho de una forma semi-automatizada, contribuyendo a mejorar la calidad de las investigaciones en los delitos que se desee investigar o encontrar un patrón detrás de la comisión de estos.

La promesa actual de los algoritmos de procesamiento de lenguaje natural es que son capaces de capturar la semántica en los documentos. En caso de no cumplirse dicha promesa, nuestros resultados serán deficientes o basados en una excesiva calibración de métricas que sólo serán válidos para un conjunto acotado de resultados experimentales.

La parte central de la hipótesis es que la vinculación de las causas puede ser realizada de forma razonable por los algoritmos, ya que es empíricamente demostrable que los humanos pueden leer dos relatos de diferentes delitos y encontrar similitudes (o diferencias) entre estos.

1.4. Metodología

El problema de analizar y evaluar la similitud entre diferentes crímenes como un mecanismo para encontrar patrones o un autor común, ha dado origen a las investigaciones más sorprendentes, como las desarrolladas por el agente del FBI John E. Douglas en la década de los 40's en su intento por desarrollar lo que hoy conocemos como *profiling*. En esta técnica, se analiza la evidencia disponible en una escena del crimen para identificar la personalidad y características conductuales del posible autor (Douglas et. al, 1986)[6].

Sin embargo, en la medida que aumenta la cantidad de registros (en este caso delitos), se vuelve imposible la labor de poder atender a todas las denuncias e indagar en posibles sospechosos o siquiera acumular evidencia concluyente que permita vincular crímenes. Es por esto que se utilizará como referencia una metodologías para la búsqueda de patrones en vastos repositorios de información: CRISP-DM (Chapman et. al, 2000).

[6]Quienes estén interesados en conocer un poco más sobre esta materia de una forma lúdica, Netflix produjo dos series en esta línea: Mindhunter (2017) y Manhunt: Unabomber (2017).

1.4.1. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) es un proceso que proporciona una sucesión de tareas para automatizar el proceso de análisis de datos y la selección de hipótesis. Una de las ventajas de CRISP-DM es que propone la existencia de un demandante que requiere de la comprensión de sus necesidades y su industrias, así como el hecho de que el proyecto no acaba una vez se encuentra el modelo idóneo, ya que después se requiere un despliegue y un mantenimiento.

En CRISP-DM la secuencia de las fases no es lineal: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase hay que realizar a continuación. Las flechas representan las dependencias más importantes y frecuentes.

Así, las 6 etapas del proceso CRISP-DM son las siguientes (Chapman et. al, 2000):

1. Entendimiento del negocio: comprender los objetivos no técnicos y los requisitos comerciales.
2. Comprensión los datos: explorar datos y comprenderlos teniendo en cuenta los objetivos comerciales.
3. Preparación de datos: limpieza, clasificación y estructura de los datos.
4. Modelamiento: aplicación y ajuste de modelos de minería de datos (o machine learning)
5. Evaluación: evaluar los resultados y revisar los pasos previos que dieron origen a los resultados que se han obtenido.
6. Despliegue: presentar y organizar el conocimiento adquirido de una manera que el cliente pueda usarlo. Esta fase puede ser tan simple como generar un informe o tan complejo como implementar un proceso de explotación de información que atraviese a toda la organización.

Dado de CRISP-DM considera de forma explícita la incorporación de conocimiento del ámbito en el que se desarrolla un proyecto, es que se considerará este proceso como la metodología principal. Así, luego de una profunda revisión bibliográfica del estado del arte de las técnicas de Text Mining, es que se desarrollan cada una de las etapas como se describe a continuación:

1. **Entendimiento del contexto**[7]: el siguiente capítulo (capítulo 2) es un esfuerzo enfocado principalmente por comprender de una forma empírica y teórica que es un delito, cuales son los móviles de su comisión y porque es importante la investigación de estos y su aprehensión. Además en el capítulo de desarrollo metodológico se profundiza en el funcionamiento del proceso de persecución penal en Chile, lo actores que forman parte de éste y las cifras Reportadas por el Ministerio público sobre los delitos que se investigan en Chile.

2. **Comprensión los datos**: en el capítulo de desarrollo metodológico se presentan las cifras macros de delincuencia para comprender cuales son los delitos de mayor connota-

[7]Es claro que no podemos referirnos a la delincuencia como un negocio, ya que es más bien un fenómeno social, cuya comprensión es en extremo compleja.

ción en Chile. Por otro lado, en el procesamiento de los datos se muestra un ejemplo de las denuncias que son recibidas por el Ministerio Público, con las particularidades que implica lidiar con texto y su falta de estructura. Finalmente, se plantea en los trabajos futuros una forma de optimizar el procesamiento de la información aprovechando la estructura narrativa o planteamiento de las denuncias.

3. **Preparación de los datos**: este es probablemente el único punto estándar de este trabajo. El preprocesamiento de los textos para la minería de estos es un proceso bastante estudiado y que se presenta en detalle en el capítulo de conceptos y definiciones (capítulo 3).

4. **Modelamiento**: uno de los principales objetivos en este trabajo tiene relación con la optimización de los parámetros de similitud para reportar o generar conjuntos de causas homogéneas en su información y contenido. En el capítulo de resultados (capítulo 5) se muestra cómo varían los resultados según cómo se escoge el parámetro de la similitud entre las causas o queries.

5. **Evaluación de resultados**: en este caso, dados los alcances del trabajo, no es posible desplegar el sistema en el Ministerio Público, en cambio, se evaluarán los resultados con especialistas en la investigación de delitos de la Fiscalía. En este caso se dara seguimiento a un foco de investigación, intentando agregar causas al foco a través de una recomendación basada en métricas de similitud semántica textual de causas en el foco con causas que no están incluidas en el foco.

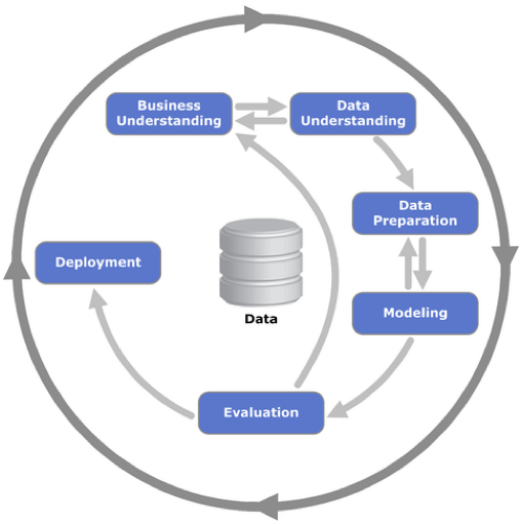


Figura 1.1: Esquema del proceso CRISP-DM

1.5. Resultados esperados

Al final de este trabajo se espera haber concretado las siguientes contribuciones:

- La implementación de los modelos en el estado del arte para el cómputo de similitud entre documentos, así como la obtención de conclusiones respecto al valor óptimo sobre parámetros de estos modelos para la búsqueda y análisis de delitos que posean información relevante para su investigación conjunta.
- El desarrollo de visualizaciones que permitan comprender los resultados de los algoritmos de búsqueda y asociación de causas.

1.6. Alcance

Acerca del desarrollo

- El presente trabajo involucra la comprehensión de la actividad y labor que desempeña el Ministerio Público, los análisis y procesamiento de las causas que ellos poseen en forma de texto, los experimentos con diferentes modelos de text mining y el desarrollo de un sistema capaz de vincular causas a través de la similitud semántica textual de las causas, además el sistema debe tener la capacidad de visualizar los resultados de estos modelos.

Sobre su uso e implementación

- El objetivo de este trabajo es realizar una prueba de concepto bajo la hipótesis de que es posible realizar una recomendación sobre qué delitos deberían ser investigados de forma conjunta basado única y exclusivamente en la información del relato del hecho y los términos de búsqueda de un analista. Por tanto, el uso, implementación y evaluación final del sistema desarrollado en este trabajo será de absoluta responsabilidad del Ministerio Público.

1.7. Estructura del documento

El resto del documento tiene la siguiente estructura: el capítulo 2 es una muy resumida presentación de los conceptos primordiales para entender la relevancia de este trabajo y la labor de la Fiscalía. Presenta una interpretación económica clásica de cómo se entiende el fenómeno de la delincuencia, la comprensión sociológica de lo que significa y cómo surge el comportamiento antisocial, presenta una interpretación jurídica de porque se debe castigar a quienes cometen un comportamiento antisocial y finalmente vincula el mundo delictual y el uso de la tecnología (contexto en que el que sitúa este trabajo).

En el capítulo 3 se presenta el marco teórico que entrega todas las definiciones para entender los conceptos y las técnicas utilizadas en el procesamiento de los documentos y el proceso de vinculación de causas. Se abordan modelos de aprendizaje de máquinas y conceptos para el procesamiento de texto.

En el capítulo 4 se presenta el desarrollo metodológico del trabajo, es decir, se detallan

las partes del proceso que darán cumplimiento al objetivo general y se especifica cómo se implementan cada uno de los modelos y herramientas de Text Mining.

En el capítulo 5 se muestran los resultados del proceso presentado en el capítulo 4, incluyendo visualizaciones y resultados estadísticos.

Finalmente, en el último capítulo se presentan las conclusiones de este trabajo a través de un análisis de sus aciertos y desaciertos. Además se incluye un subcapítulo en las conclusiones que da cuenta de algunos trabajos futuros.

Capítulo 2

Comprensión del delito, el actuar criminal y su aprehensión

En la sociedad anómica existe un desajuste entre fines proyectados sobre los ciudadanos y medios lícitos para alcanzar dichos objetivos, generándose gravosas fuentes de presión anómica, que en determinados casos podrán desencadenar comportamientos delictivos y antisociales para lograr alcanzar el 'éxito' monetario y social.

Robert K. Merton, 1938

2.1. Interpretación económica del delito

Las teorías del comportamiento criminal basadas en el supuesto de una elección racional fueron inicialmente propuestas por autores de la llamada Escuela Clásica de la Criminología, principalmente por los autores Beccaria (1764) y Bentham (1843). Bentham señala que "el beneficio del crimen es la fuerza que impulsa al hombre a la delincuencia: el dolor del castigo es la fuerza empleada para retenerlo. Si la primera de estas fuerzas es mayor, el crimen se cometerá; si es el segundo, el crimen no se cometerá".

La propuesta de Bentham se revitalizó y modernizó en el artículo *Crime and Punishmen* de Becker (1968), quién se basa en la economía moderna de bienestar y señala que: "una persona comete una ofensa si su utilidad esperada por el acto excede la utilidad que podría obtener mediante el uso de su tiempo y otros recursos en otras actividades. Algunas personas se convierten en *delincuentes*, por tanto, no porque su motivación básica difiera de la de otras personas, sino porque sus beneficios y costos son diferentes". Así mismo, el autor agrega, "el comportamiento delictivo se convierte en parte de una teoría general y no requiere conceptos ad hoc de asociación diferencial, comportamiento desviado o similares". El enfoque económico de Becker no supone que los seres humanos necesariamente tengan información completa ni que sean conscientes de sus esfuerzos por maximizar su utilidad.

El modelo de Becker plantea una función de utilidad esperada para cada delito (U_j), en donde un factor central lo juega la probabilidad de ser condenado por el delito cometido (p_j).

Esta probabilidad de condena pondera la utilidad de cometer un delito y no ser capturado, así como la utilidad de cometer el delito y ser capturado, lo cual implica un beneficio Y_j, pero un costo de f_j, que es el castigo en caso de ser condenado. Lo anterior se resume en la ecuación 2.1.

$$\mathbb{E}[U_j] = p_j U_j(Y_j - f_j) + (1 - p_j)U_j(Y_j) \tag{2.1}$$

Es necesario señalar que el trabajo original de Becker no sólo introduce una ecuación que permite vislumbrar desde la teoría del bienestar cómo modelar el crimen, sino que también postula (según la misma óptica), cuál es el costo social del castigo, cúal es el óptimo de los recursos policiales, el costo de aprehensión y castigo de los criminales, los daños sociales netos producidos por los delincuentes y cuál será el nivel de delitos observados dada una determinada jurisprudencia. Por otro lado, el modelo de Becker es destacado por ser uno de los trabajos germinales de la economía del delito, pero en ningún caso representa de forma figurativa el entendimiento actual sobre la materia.

Finalmente, es directo señalar que bajo el supuesto de que este modelo explica porque un delincuente decide cometer un crimen, la forma de disminuir los niveles de criminalidad sería aumentar la probabilidad de ser condenado por un delito. Lo anterior se debe a que incluso si se aumentan mucho las penas, esto sólo se verá reflejado en la utilidad resultando si es que el delincuente es capaz de internalizar dicha información en su propia función de utilidad. No obstante, décadas de investigación sobre el efecto de la severidad de la penas o la certidumbre de ser castigado se inclinan por relevar el efecto de la certidumbre en ser castigado. Robinson y Darley (2004), resumen un centenar de trabajos científicos señalando que los potenciales delincuentes a menudo no conocen las reglas legales. Incluso si lo hacen, con frecuencia no pueden utilizar este conocimiento para guiar su conducta, debido a una variedad de factores situacionales, sociales o químicos. Incluso si es posible, un análisis racional generalmente hace que los beneficios percibidos del delito sean mayores que los costos percibidos, debido a la variedad de realidades de la justicia penal, como las bajas tasas de castigo. Estas conclusiones se ven reforzadas por estudios de índices de criminalidad luego de cambios en las reglas del sistema penal respecto a las condenas, donde muchos no demuestran un efecto disuasivo.

2.2. La criminología y el comportamiento antisocial[1]

En primer lugar, definiremos la criminología como la ciencia que estudia el comportamiento delictivo y antisocial en sus dimensiones real y percibida, y los mecanismos de control social formal e informal empleados para la prevención, control y tratamiento de la criminalidad, el infractor y la víctima, con el fin último de velar por el bienestar personal y social del conjunto de la ciudadanía. Es necesario matizar que el estudio de la Criminología sobrepasa los límites legales fijados por la regulación penal, estudiando también comportamientos me-

[1]En esta sección se presenta de forma resumida algunos de los hitos más relevantes de la criminología, basado en el excelente trabajo realizado por David Buil Gil (2016) titulado *¿Qué es la criminología? Una aproximación a su ontología, función y desarrollo.*

ramente desviados o antisociales que, de acuerdo con una definición estrictamente jurídica del término, no podrían ser catalogados como delitos (Gil, 2016).

Los orígenes de la criminología se remontan a la obra del ilustrado italiano Cesare Beccaria, en su libro *Dei delitti e delle penne (1764)*, quien sienta las bases de la Criminología empírica al señalar que la sociedad debe estudiar científicamente los delitos y los medios para su prevención. Este autor, es uno de los primeros pensadores en señalar que el fin de las penas es la prevención de las infracciones para la protección del orden social, planteando una reforma necesaria en el sistema de justicia penal para hacerlo más humano y justo (Cid y Larrauri, 2001).

Más tarde, algunos autores de la Escuela Cartográfica fueron los primeros en buscar estadísticas sobre los crímenes, con lo que encontraron algunas relaciones como la mayor propensión al delito entre varones jóvenes, la tendencia a los delitos violentos en invierno y a los delitos contra la propiedad en verano, y la correlación entre heterogeneidad étnica, marginalidad y tasas delictivas superiores (Hagan, 2010). En el periodo de esta escuela, Quételet (1831) y de Candolle (1830) constataron lo que posteriormente seria denominado la cifra negra, que constituye aquellos sucesos que por determinados criterios pueden ser considerados delitos, pero que no quedan registrados por las fuentes de datos encargadas de registrar la delincuencia" (Biderman y Reiss, 1967).

Posteriormente, y en conjunto con la aparición de la Escuela Positiva, aparece la obra *L'uomo delinquente* (1876), escrita por el médico italiano Cesare Lombroso, quien estudió la estructura anatómica y los cráneos de muestras de delincuentes condenados, extrayendo de ello una serie de especificidades físicas que caracterizaban a los sujetos desviados: frente baja y salida, pómulos supradesarrollados, asimetrías y poca capacidad craneal, dimensión anormal de las orejas, entre otros. Sin embargo, no encontraron apoyo empírico en los estudios desarrollados por los seguidores de Lombroso, por no basarse en metodologías rigurosas y sistematizadas (Garrido et al, 2006).

La última de las escuelas, pero no por eso necesariamente la actual, es la denominada Escuela de Chicago. Esta escuela, trata de estudiar cómo los cambios en las estructuras de organización social en las grandes ciudades de principios del siglo XX se relacionan con las causas de la desviación (Cullen y Agnew, 2011). Uno de los principales trabajos es el de Shawk y McKay (1942), que a partir del estudio de la distribución de la criminalidad juvenil en las diferentes zonas de la ciudad de Chicago, elaboran la teoría de la desorganización social, a partir de la cual explican que existen determinados factores ecológicos, entre los que destacan la pobreza, la movilidad, la multiculturalidad, o la degradación física del espacio urbano, localizados en mayor medida en unos barrios que en otros, que se relacionan con una menor capacidad de las comunidades para ejercer control sobre la desorganización social, elemento que permite explicar la diferencia en las tasas de delincuencia en las diferentes zonas de la urbe (Cid y Larrauri, 2001).

Luego de estas escuelas, han aparecido una gran cantidad de teorías que han sido ampliamente aceptadas hasta la actualidad cómo la teoría de la asociación diferencial, las teorías de la anomia y la tensión, las teorías del control y las teorías de las oportunidades (estrechamente relacionada con el triángulo del delito 2.2.1). Muchas de las anteriores son integradas en los actuales enfoques teóricos integradores, cómo la teoría integrada del potencial cognitivo

antisocial (ICAP) de David Farrington (2017) y la teoría del triple riesgo delictivo (TRD) de Santiago Redondo Illescas (2015).

A pesar de la extensa literatura existente en criminología y su madurez, pocos enfoques hacen referencia a cómo caracterizar un delito, lo cual posee una importancia fundamental desde el punto de vista de las policías, ya que en caso de buscar la reducción de un determinado tipo de crimen, es necesario en primer lugar caracterizar el tipo de delito, su autor y la víctima.

2.2.1. Triángulo del delito

Este es una formalización y extensión de la teoría de crimen predatorio (el término original es *routine activity approach*) (Cohen y Felson, 1979), es decir, cuando la actividad criminal constituye parte de la rutina de un individuo. Estos crímes son generalmente aquellos en donde converge una victima indefensa, un objetivo alcanzable y la ausencia de un guardián que disuada o impida el acto criminal (Clarke y Eck, 2014)[2].

Del triángulo del delito (Ver Figura 2.1) se deriva la clasificación de los problemas de *recurrencia* más comunes:

1. Criminal prolifero: el concepto hace referencia al actuar reiterado de un criminal actuando en diferentes lugares o de forma reiterada sobre diferentes objetivos o víctimas. Un delincuente armado que busca asaltar personas de noche es un ejemplo de este tipo de problemas.

2. Víctimas reiteradas: este término hace referencia a aquellas personas son atacadas de forma reiterada por diferentes delincuentes. Taxistas asaltados por múltiples delincuentes en diferentes ocasiones son un ejemplo de este problema.

3. Lugar reiterado: también conocido como *hot spot*, involucra la interacción diferentes criminales y víctimas interactuando en el mismo lugar. Una calle donde usualmente se roban vehículos durante la noche es un ejemplo de un *hot spot*.

Comprender cómo surgen estos problemas recurrentes han contribuido a pensar en lo que se podría hacer no solo para arrestar a los delincuentes, sino también para evitar que vuelvan a delinquir haciendo un mejor uso de los manipuladores[3]; qué pueden hacer las víctimas para reducir la probabilidad de ser objeto de un delincuente y qué cambios podrían hacerse a los lugares donde ocurren delitos de forma reiterada, ya sean estas escuelas, medios de transporte público, estacionamientos, etc.

[2]La combinación de delincuente, objetivo u víctima y ausencia de guardianes que impidan o disuadan el delito sería descrita más tarde como la *química del crimen* (Felson, 1994).

[3]Los manipuladores generalmente son personas que poseen algún tipo de poder o control sobre el delincuente y que lo pueden hacer delinquir nuevamente, así como también pueden contribuir a rehabilitar al sujeto y/o impedir que vuelva a cometer actos criminales.

Figura 2.1: Triángulo del delito

2.3. Interpretación jurídica del delito y su castigo

En su acepción etimológica, la palabra delito deriva del verbo latino *delinquere*, que significa abandonar, apartarse del buen camino, alejarse del sendero señalado por la ley. En este caso, abandonar la ley (De Pina, 2003).

Francisco Carrara (1991), define al delito como la infracción de la ley del Estado, promulgada para proteger la seguridad de los ciudadanos, y que resulta de un acto externo del hombre, positivo o negativo, moralmente imputable y políticamente dañoso.

El Código Penal de Chile[4], en su Artículo 1 señala: "Es delito toda acción u omisión voluntaria penada por la ley. Las acciones u omisiones penadas por la ley se reputan siempre voluntarias, a no ser que conste lo contrario". En el mismo artículo, se señala respecto al autor: "El que cometiere delito será responsable de él e incurrirá en la pena que la ley señale (..)". Además, respecto a la disimilitud entre delito y crimen, el mismo documento en su Artículo 3 señala: "Los delitos, atendida su gravedad, se dividen en crímenes, simples delitos y faltas y se califican de tales según la pena que les está asignada en la escala general del art. 21". Por tanto, los delitos y los crímenes son definidos por los distintos ordenamientos jurídicos vigentes en un determinado territorio e intervalo de tiempo.

2.3.1. La importancia de la pena y su función

En primer lugar, es importante señalar que así como las ciencias naturales persiguen establecer pautas de orientación con respecto a la naturaleza, el Derecho tiene como función establecer pautas de orientación con respecto a los integrantes del sistema social (Jakobs, 2006), y es precisamente el orden de este sistema social, el que la pena pretende mantener.

[4]Existen diferencias en la definición de delito entre el Código Penal y el Código Civil de nuestro país. Por simplicidad y cercanía con el contexto, se ha optado por recoger la definición del Código Penal.

Günther Jakobs, basado en el pensamiento de Niklas Luhmann[5], comprende que son las expectativas los elementos estructurales del sistema, esto en la medida de que todo orden social se basa en la existencia de ciertas expectativas de comportamiento más o menos estables. La confianza en dichas expectativas, garantizadas por las sanciones, son un mecanismo de reducción de la complejidad social.

En caso de existir una defraudación de las expectativas (del comportamiento de los integrantes del sistema social), esto provoca que la sociedad abandone dichas expectativa, generando un aprendizaje y un cambio en la visión que se tiene del mundo. Así, la pena, es concebida como un instrumento para resolver las defraudaciones de expectativas que no pueden ser estabilizadas de otra manera, evitando este aprendizaje y cambio en la visión que se tiene del mundo, con lo cuál las personas pueden mantener firme su confianza en las mismas, a pesar de la defraudación.

2.4. La vinculación criminal y el uso de la tecnología

Según algunos estudios, una gran proporción de los delitos son cometidos por una minoría de delincuentes (Borg et al. 2014; Tonkin et al. 2012). Por ejemplo, en los Estados Unidos, se encontró que el 5 % de los delincuentes estaban involucrados en el 30 % de los delitos (Tonkin et al. 2012). En Chile, Peillard (2013) en conjunto con Paz Ciudadana, han estimado que cerca del 65 % de los criminales son reincidentes y más del 50 % de aquellos que han sido sentenciados por los tribunales, vuelven a los recintos penales imputados por el mismo tipo de delito por el cual fueron condenados de forma previa.

Por tanto, vincular delitos que compartan un *modus operandi (MO)* es de gran importancia para la aplicación de la ley por varias razones. En primer lugar, la agregación de información de diferentes delitos aumenta la cantidad de evidencia disponible. Por ejemplo, el somatotipo del delincuente visto en un caso y el acento del delincuente escuchado en otro caso, podrían usarse de forma conjunta para perfilar a un delincuente si estos dos casos se logran identificar como causas con un autor común. En segundo lugar, la investigación conjunta de crímenes múltiples permite un uso más eficiente de los recursos policiales (Woodhams et al, 2007).

Los sistemas de vinculación criminal se remontan al desarrollo del programa de aprehensión de crímenes violentos del FBI en 1985[6]. ViCAP fue desarrollado con el propósito loable de evitar una "ceguera de vínculos", un término utilizado para describir la ausencia de comunicación entre los organismos encargados de hacer cumplir la ley que podrían estar investigando casos relacionados (Egger, 1984). El programa ViCAP buscó reducir este problema ayudando a las agencias a determinar si delitos vinculados se estaban cometiendo o no entre fronteras jurisdiccionales. Para lograr este objetivo, la información sobre delitos violentos se ingresó en una base de datos informática y se analizó para identificar delitos que mostraban distin-

[5]Niklas Luhmann fue un sociólogo alemán reconocido por su formulación de la teoría general de los sistemas sociales.

[6]Si bien el ViCAP es el primer sistema de información conocido en este ámbito, la *Royal Canadian Mounted Police* desarrolló el ViCLAS, plataforma que es utilizada en Alemania, Australia, Austria, Bélgica, Holanda, Francia, Irlanda, Nueva Zelanda, República Checa, Reino Unido, Suiza y dos estados de Estados Unidos.

tos patrones de similitud que podrían reflejar vínculos. Hasta el día de hoy, ViCAP sigue siendo un sistema nacional de información de datos para recopilar, clasificar y analizar casos resueltos y no resueltos de delitos violentos (Bennell et al, 2012).

No obstante, incluso con los nuevos enfoques basados en el procesamiento de datos para la detección y/o predicción de delitos, el trabajo fundamental de los analistas delictivos sigue siendo difícil, y a menudo, manual; patrones específicos de delincuencia no son necesariamente fáciles de encontrar por medio de herramientas tecnológicas. El método más frecuente (y el más exitoso) para identificar patrones de delitos específicos implica la revisión de informes de delitos diariamente y la comparación de esos informes con delitos pasados (Gwinn et al, 2013). El problema es que dicha labor es intensiva en el uso de analistas y tiempo de análisis de los casos. Al hacer estas comparaciones, un analista busca elementos en común entre un delito pasado y uno presente para sugerir un patrón. Por tanto, las herramientas para resolver este problema podrían ser extremadamente valiosas para ayudar a los analistas y podrían conducir directamente a medidas preventivas. Detectar estos patrones automáticamente es un desafío en que las herramientas de aprendizaje automático y la minería de datos pueden contribuir directamente al trabajo de los analistas de crímenes y delitos.

2.4.1. Modelos de vinculación criminal

La oportunidad es clara y la necesidad latente por desarrollar modelos y sistemas que permitan aumentar la capacidad de análisis en la investigación de delitos sin resolver. No obstante, pese al tamaño de la oportunidad, sólo una decena de autores se han propuesta desarrollar modelos para la vinculación criminal. Lo anterior se hace patente con el hecho de que recién el año 2002 se acuña el término de vinculación criminal (crime linkage) en la literatura con el trabajo de Benell y Canter (2002). Luego de esto han aparecido una serie de trabajos como el Bennell y Jones (2005), Woodhams y Toye (2007), Tonkin y Grant (2008), Snook et al (2012), Bennell et al (2012), Reich y Porter (2015), Porter (2016), Chi et. al (2017). Sin embargo, cada uno de los trabajos anteriores, desarrollan modelos sobre bases de datos parametrizadas, basadas en clustering de causas o clasificación de multiclases. Lo anterior no es replicable en Chile, ya que el 100 % de las causas es recibida mediante datos no estructurado, en este caso, texto transcrito por un funcionario que recibe una denuncia. Dado lo anterior, en este trabajo se abordará el problema de la vinculación criminal, pero desde una perspectiva totalmente nueva y experimental, ya que se desarrollará un modelo de vinculación criminal basado en modelos que reciban cómo input el texto de las denuncia en poder del Ministerio Público.

Capítulo 3

Modelos y conceptos de text mining

3.1. Conceptos básicos

3.1.1. Preprocesamiento de texto

El preprocesamiento es uno de los componentes claves en cualquier proceso de minería de datos (o texto en este caso), pues permite eliminar del texto el ruido existente o caracteres que no contribuyen con información relevante. Mediante análisis experimentales, Uysal y Gunal (2014) revelan que elegir combinaciones apropiadas de tareas de preprocesamiento puede proporcionar una mejora significativa en la precisión de tareas clasificación de texto. Esta tarea típicamente involucra 4 etapas: Tokenización, filtrado, Lematización y Stemming.

Stop words y eliminación de otros caracteres

Wilbur y Sirotkin (1992) definen el concepto de stop word como: "una palabra que tiene la misma probabilidad de aparecer tanto en aquellos documentos relevantes para una consulta como para aquellos que no". Por tanto, la remoción de estas palabras implican una simplificación en las labores de text mining al eliminar contenido que no aporta información. En este trabajo, se removieron las stop words mediante el uso de la librería NLTK[1] de Python.

Así mismo, números y caracteres especiales sólo contribuyen a añadir ruido en el texto, por tanto su remoción fue seguida de la remoción de stop words utilizando expresiones regulares[2] en Python.

[1]NLTK: http://www.nltk.org/

[2]Un extracto de la definición de Wikipedia para expresiones regulares es: "En el área de la programación las expresiones regulares son un método por medio del cual se pueden realizar búsquedas dentro de cadenas de caracteres. Sin importar la amplitud de la búsqueda requerida de un patrón definido de caracteres, las expresiones regulares proporcionan una solución práctica al problema".

Lematización y Stemming

Allahyari et al (2017) define la lematización cómo: "la tarea que considera el análisis morfológico de las palabras, es decir, agrupa las diversas formas de una palabra para que puedan analizarse como un solo elemento. Dicho de otra forma, los métodos de lematización intentan mapear las formas verbales a tiempo infinito y sustantivos en una sola forma". Por ejemplo: $\{bonito, hermoso, lindo\} \rightarrow \{bello\}$.

Los mismos autores señalan respecto al stemming: "los métodos de stemming apuntan a obtener la raíz de las palabras. Los algoritmos de Stemming son de hecho dependientes del lenguaje". Por ejemplo: $\{escrito, escribi, escribiendo\} \rightarrow \{escribir\}$.

Ambos métodos (lematización y stemming) buscan reducir el abanico de palabras que significan lo mismo y permiten realizar una comparación más sencilla entre los documentos.

3.1.2. Term-Document Matrix

Con el objetivo de representar una colección de textos se crea el Term-Document Matrix. Las filas de la matriz son palabras individuales y las columnas son documentos o unidades más pequeñas, como frases o extractos de un documento, según corresponda para cada aplicación. Las entradas de celda individuales contienen la frecuencia con la que un término ocurre en un documento, por tanto la celda (i, j) representa la frecuencia del término i en el documento j. Una importante consideración de esta matriz es que el orden de las palabras en el documento no es importante en esta representación (a menudo el Term-Document Matrix es conocido como *bag-of-words*).

	Doc 1	Doc 2	Doc 3
Term 1	4	0	1
Term 2	0	1	1
Term 3	0	2	0
Term 4	1	3	1

Figura 3.1: Representación del Term-Document Matrix

3.1.3. Neural Networks

En 1943, el neurofisiólogo Warren McCulloch y el matemático Walter Pitts escribieron un artículo sobre cómo podrían funcionar las neuronas en el cerebro. Para esto, modelaron una red neuronal simple usando circuitos eléctricos. Este trabajo es conocido como el origen de este algoritmo y es por lo que comunmente se señala que las redes neuronales artificiales

poseen una inspiración biológica[3].

Más tarde, en 1958 Frank Rosenblatt implementó la idea del perceptrón y mostró que podría usarse para aprender a clasificar formas simples correctamente con entradas de 20x20 pixeles[4]. El perceptrón es la unidad básica de las redes neuronales y simulan poseer la misma función y fisionomía de las neuronas (Ver Figura 3.2). El perceptrón recibe datos parametrizados como entrada, los cuales posteriormente pondera mediante la suma de los productos punto entre la información de la entrada y los pesos; este valor (real) es evaluado posteriormente en una función de activación. Las funciones de activación replican la *ley del todo o nada*[5] que gobierna la actividad neuronal. Esto se traduce –de forma general–, en que una función de activación asignará un valor real (digamos α) si es que la suma ponderada de pesos y entradas alcance un umbral (lo designaremos γ), o de lo contrario asignará el valor 0. Esto se traduce formalmente en lo expresado en la ecuación 3.1.

$$\begin{aligned} Y &= f\left(\sum_{\mathrm{i}=1}^{N} w_{\mathrm{i}} x_{\mathrm{i}} + \theta\right) \\ f(x) &= \begin{cases} \alpha, & \text{si } x \geqslant \gamma \\ 0, & \text{de lo contrario} \end{cases} \end{aligned} \tag{3.1}$$

Las funciones de activación generalmente se seleccionan dentro del conjunto funciones que han demostrado ser efectivas en alguna tarea en particular y que cumplen con la labor designada de entregar valores en un cierto rango si se alcanza un umbral o entregar el valor 0 de lo contrario. Algunas de las funciones más utilizadas son:

- ReLu: $R(x) = max(0, x)$
- Sigmoide: $\theta(x) = \frac{1}{1+\mathrm{e}^{-x}}$
- Tangente hiperbólica: $tanh(x) = \frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}$

Las redes neuronales son la representación de múltiples perceptrones ordenados de forma paralela en lo que se denominan capas, donde el resultado (o salida) de una capa alimenta como entrada a la otra (Ver Figura 3.3). La arquitectura más clásica de redes neuronales es conocida como *Feedforward* o *Multi-Layer Perceptron (MLP)*. En redes MLP, la primera capa se llama capa de entrada, es donde se introduce un punto de datos en forma vectorial, la última capa se llama capa de salida y todas las capas intermedias se llaman capas ocultas.

[3]Para quienes esten interesados en conocer en detalle la historia de las redes neuronales, recomiendo vistar las partes 1, 2, 3 y 4 que Andrey Kurenkov escribio en su *blog* A 'Brief' History of Neural Nets and Deep Learning.

[4]Dada la popularidad de los algoritmos de Deep Learning 3.1.4, probablemente en unos años, si alguien revisita la historia de las redes neuronales, podrá aseverar que con este hito nació el aprendizaje de máquinas.

[5]La ley del todo o nada señala que si un estímulo es de la intensidad suficiente como para desencadenar un impulso nervioso, este impulso se traduce de forma íntegra. De lo contrario, si el estímulo es débil, este no se traduce, por lo que no producirá una reacción débil.

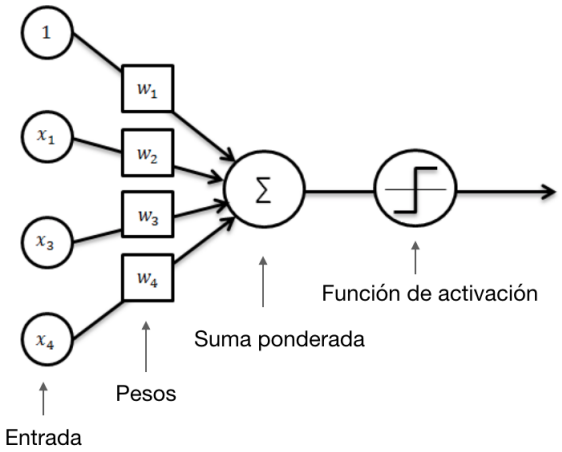


Figura 3.2: Representación del Perceptrón

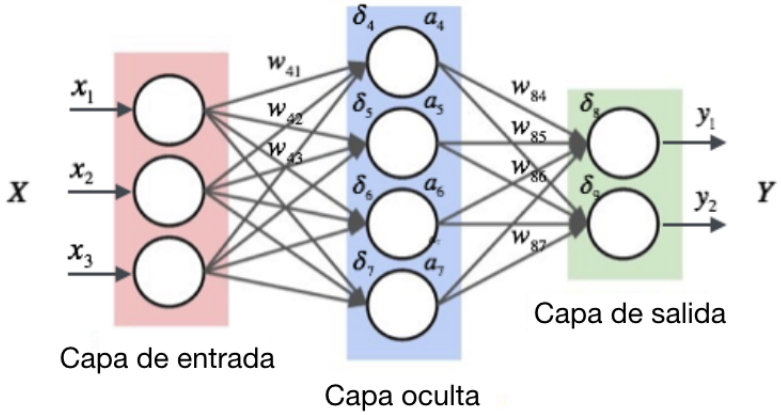


Figura 3.3: Representación de una red Multi-Layer Perceptron

3.1.4. Deep Learning

Está demostrado que una red MLP con una capa oculta lo suficientemente grande puede ser entrenado para emular cualquier función posible. En la práctica, sin embargo, es preferible agregar más capas ocultas consecutivas, lo cual da origen al término *Deep learning*, que consiste básicamente en una red neuronal con muchas capas ocultas.

A pesar del gran éxito de las técnicas de aprendizaje basado en deep learning, todavía no existe una comprensión teórica *profunda* de porque estas arquitecturas de redes funcionan tan bien. Probablemente la mejor explicación hasta el momento ha sido realizada por el trabajo de Shwartz-Ziv y Tishby (2017), quienes –y resumido en de la forma más sencilla posible– muestran que una mayor cantidad capas permiten descomponer el complejo problema de maximizar la información mutua[6] entre la capa de entrada y la de salida $I(X;Y)$ entre las múltiples capas de la red. Esto implica que un mayor número de capas permiten construir

[6]La información mutua es una medida que cuantifica la cantidad de información que puede ser explicada de una variable aleatoria a partir de otra. Cuando la información mutua es cero, significa que las variables son independientes.

una mejor y más flexible representación de los datos. Así, un mayor número de capas acelera el proceso de entrenamiento debido a la mayor flexibilidad adquirida conforme aumentan el número capas ocultas, al requerir un menor número de épocas[7] para completar la tarea de entrenamiento. Además cada capa extra contribuye a crear un mejor representación de los datos y mejorar los resultados del algoritmo.

Las redes neuronales profundas proveen de muy buenos resultados en la mayoría de las tareas en las que han sido testeadas, sin embargo, estas demandan un uso intensivo de hardware y de grandes cantidades de datos para ser entrenadas de forma apropiada. También es necesario entrenar este algoritmo con métodos de regularización como *drop-out*[8] para evitar el sobreajuste a los datos, ya que como ha sido mencionado anteriormente , las redes neuronales profundas proveen la oportunidad de general una representación muy flexible de los datos, lo que eventualmente significar el ajuste perfecto a los datos de entrenamiento y la pérdida de generalidad. Sin embargo, lo que realmente ha vuelto tan popular a las redes neuronales profundas en los últimos años son 2 tipos de arquitecturas que han permitido empujar la barrera del conocimiento y el espacio de oportunidades en el campo de aplicaciones del aprendizaje de máquinas.

- Redes Neuronales Convolucionales (CNN): en solo unos pocos años, las CNN han empujado el estado del arte en la mayoría de las tareas de visión computacional. La forma tradicional de trabajar con imágenes consistía simplemente en considerar una imagen como un vector de entrada (donde cada pixel es una observación) e introducir este vector en una gran red MLP. Esto, en la práctica, es inviable y extremadamente intensivo en hardware. Incluso con los recursos computacionales modernos, entrenar un modelo con tanta alta dimensionalidad es sumamente complejo.

  Un filtro (o *kernel*) en el contexto del procesamiento de imágenes es una matriz de ponderadores. La idea general es que los filtros se pasan sobre una imagen para transformar estos datos, es decir, en vez de considerar sólo un peso por registro de entrada como en las redes clásicas, se considera una matriz con estos pesos (Ver Figura 3.4). Una vez el filtro ha recorrido un conjunto de píxeles, los valores de estos píxeles se multiplican por el peso del filtro .[en]cima de ellos", luego se suman y el valor del píxel central por el cual pasa el filtro se reemplaza por esa suma.

  En una CNN, una neurona es un filtro que pasa sobre toda la imagen. Cada capa de neurona produce un conjunto de representaciones alternativas de una imagen, que luego son transformadas por capas posteriores. Así, la labor de etiquetar una imagen (por ejemplo número, letras, etc.) puede ser resuelto en un tiempo mucho menor. Así mismo, para resolver las tareas de reconocimiento, que son esencialmente una tarea de reducción de dimensionalidad, los CNN también incluyen capas de agrupamiento (*pooling*). El método de agrupamiento más común es *max pooling*, donde el tamaño de una imagen se reduce seleccionando el píxel con el mayor valor entre sus píxeles vecinos.
- Redes Neuronales Recurrentes (RNN): RNNs son la solución que se propuso para hacer frente a las secuencias temporales. Una capa RNN entrena una única celda, que es

[7]Cada época representa una iteración en la que los pesos de cada perceptrón de la red son calibrados para mejorar los resultados en la tarea para la que se está entrenando el algoritmo.

[8]La idea central detrás de esta técnica es aleatoriamente desconectar perceptrones de la red neuronal (el perceptrón y sus conexiones) durante el entrenamiento. Esto previene que estas unidades se puedan ajustar demasiado (sobre ajustar).

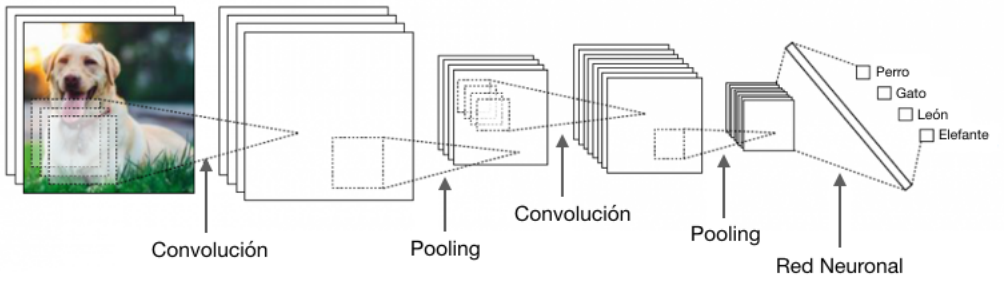


Figura 3.4: Representación de una CNN

una red neuronal con arquitectura arbitraria que toma un elemento en una secuencia representada como un vector y emite otro vector del mismo tamaño. La clave es que un vector adicional del mismo tamaño llamado estado se transfiere de una instancia de la celda a otra. Una celda procesa una secuencia de vectores de entrada en secuencia, cambia el estado y lo mantiene para procesar el siguiente elemento. Es decir, para decidir que es lo que sucederá en el siguiente estado se considera la información del estado actual y los estados anteriormente (Ver Figura 3.5).

Los RNN se usan de varias formas. Se pueden usar para leer texto y generar un solo etiquetado representado por el último estado. También se pueden usar para tomar una entrada de tamaño fijo en el primer paso y generar una oración. Pueden tomar una oración y producir otra (como en las tareas de traducción de idiomas). Finalmente y lo más relevante para este trabajo es que el estado oculto (conjunto de capas ocultas) de una RNN durante su entrenamiento son la representación vectorial que llamaremos *Word Embeddings* en lo que sigue. Las RNN han revolucionado el campo del procesamiento de lenguage natural, así como las CNN han revolucionado la visión computacional.

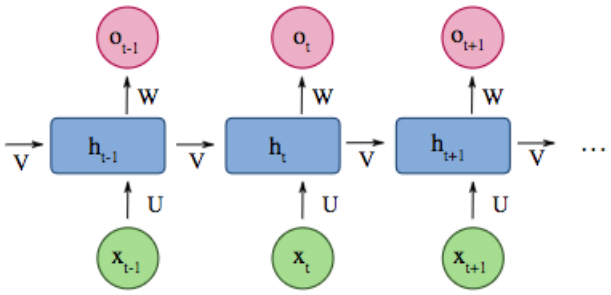


Figura 3.5: Representación de una RNN

3.2. Recuperación de documentos

Un sistema de recuperación de documentos consta básicamente de 3 partes: la representación de los documentos (pensar en la matrix term-document), la representación de los requisitos de los usuarios (comúnmente conocido como query) y los algoritmos utilizados para hacer coincidir los requisitos del usuario (query) con la representaciones de los documentos

de forma de recuperar los documentos que poseen el contenido más similar a lo estipulado en la query. En la Figura 3.6 se aprecia una esquematización del proceso.

El problema fundamental que se busca resolver con cualquier sistema recuperación de documentos es cómo recuperar sólo los documentos relevantes para los requisitos de información del usuario, sin recuperar los no relevantes. Para esto existen numerosos algoritmos de recuperación, sin embargo, en este trabajo la evaluación cuantitativa en términos de métricas de desempeño es muy difícil dada la falta de etiquetas y categorización de los datos/documentos. Es por esto que se ha optado por evaluar e implementar uno de los clásicos de la literatura.

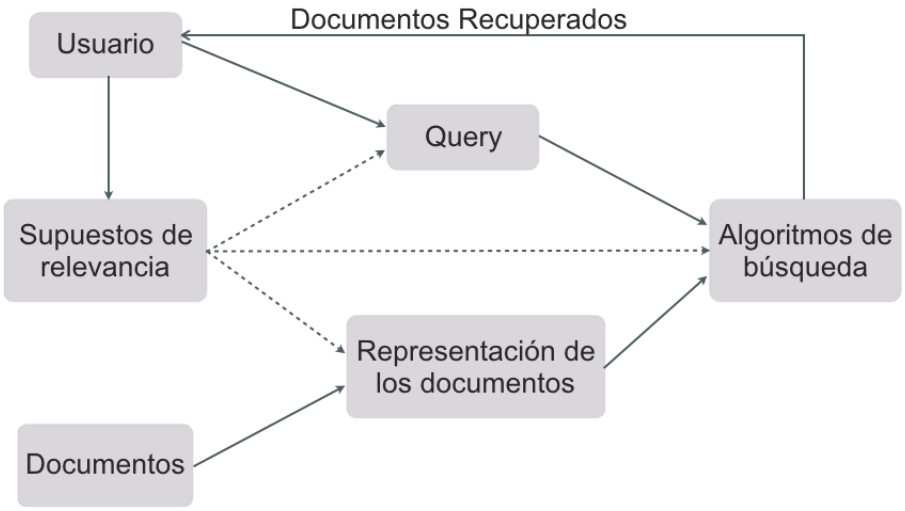


Figura 3.6: Representación de un sistema de recuperación de documentos

3.2.1. Latent Semantic Indexing

Latent semantic Indexing (LSI), (Deerwester et al., 1990) también llamado *Latent semantic Analysis* (LSA) fuera del contexto de recuperación de documentos es un modelo introducido por Dumais, Furnas, Landauer y Deerwester en 1988 a través de una patente titulada *"Computer information retrieval using latent semantic structure"*. Mas tarde en los años 1990 y 1995 se publican dos artículos científicos dando cuenta de los detalles matemáticos detrás del modelo. De forma sencilla diremos que este modelo opera de la siguiente manera (Dumais, 2004):

1. Term-Document Matrix (3.1): se utiliza como input esta representación vectorial de los documentos.
2. Transformed Term-Document Matrix: en lugar de trabajar con las frecuencias de términos, las entradas en la matriz de documento de términos a menudo se transforman. El mejor rendimiento se observa cuando las frecuencias se acumulan de forma sublineal, generalmente $Log(freq_{ij} + 1)$, e inversamente con la ocurrencia general del término en la colección.
3. Reducción de dimensionalidad: sobre la matriz calculada se realiza una descompresión de valores singulares (SVD) de rango reducido en la que se conservan los k valores

singulares más grandes y el resto se establece en 0. La representación SVD de dimensión reducida resultante es la mejor aproximación k-dimensional al matriz original, en el sentido de mínimos cuadrados.

4. Recuperación en espacio reducido: las similitudes se calculan entre entidades en el espacio de dimensión reducida, en lugar de ser calculadas en la matriz de documentos y términos original. Debido a que tanto los documentos como los términos se representan como vectores en el mismo espacio, las similitudes document-document, term-term y document-term son sencillas de calcular. La similitud coseno 3.6.1 entre vectores se usa como medida de similitud para muchas aplicaciones de recuperación de información porque ha demostrado ser efectiva en la práctica.

La explicación matemática de este modelo se basa en que se puede realizar una Singular Value Decomposition (SVD) de una matriz $D \in \mathbb{R}^{mxn}$, donde mxn representa la dimensión de term-document, lo cual suele denotarse cómo $D = U\Sigma V^T$. Luego de haber realizado la SVD, tenemos que $U \in \mathbb{R}^{mxr}$, $\Sigma \in \mathbb{R}^{rxr}$ y $V^T \in \mathbb{R}^{rxn}$, siendo r el rango de D y además, en la matrix Σ los valores singulares (singular values) se ordenan tal que $\sigma_1 \geq \sigma_2 \geq .. \geq \sigma_r > 0$. Luego, el objetivo de este modelo es aproximar la matrix original D de rango r por una matriz $\hat{D}$ de rango k con $k \ll r$, comando los primero k-valores singulares de la matriz Σ. Así, se puede demostrar que por el teorema de Eckart-Young que $\hat{D}$ es la mejor aproximación a D entre todas las matrices de rango k. Lo cual se resume cómo:

$$D \simeq \hat{D} = \hat{U}\hat{\Sigma}\hat{V}^T \tag{3.2}$$

La suposición subyacente de LSI es que los valores singulares pequeños representan ruido en el corpus, de modo que los vectores de term-document originales se ubican en demasiadas dimensiones. Esto es ciertamente plausible, ya que los términos individuales exhiben correlaciones, y es bastante probable que no todos los términos en un documento determinado sean necesarios para proporcionar una representación precisa de los mismos. Así, el objetivo es tener una mejor representación de los elementos y sus relaciones entre sí, y como los términos muestran correlaciones, pueden existir mejores vectores de documentos en un espacio k-dimensional, con $k \ll r$.

Finalmente, para recuperar los documentos que más similitud poseen con un query, se computa la similitud coseno entre la representación vectorial de la query y la representación vectorial de los documentos en este nuevo espacio k-dimensional ($cos(\hat{\mathrm{d}}_\mathrm{i}, q)$)

Se ha escogido este modelo en primer lugar por su amplia validación en la literatura, pero principalmente a su robustez frente a la sinonimia y la existencia de hiperónimos[9], lo cual constituye un problema mayor al momento de realizar búsquedas a través de queries, ya que si de desea buscar por ejemplo robos con arma y se introduce sólo la palabra arma en la query, un algoritmo de búsqueda textual sólo recuperará aquellas causas donde aparezca esa palabra, ignorando palabras claves cómo: revolver, cuchillo, pistola, etc. Nuestra suposición es que al realizarse la aproximación de D a $\hat{D}$, la representación vectorial de palabras semánticamente iguales se fusionan o posicionan es vectores muy cercanos, permitiendo recuperar palabras símiles a los términos de búsqueda.

[9]Según la RAE, hiperónimo: Palabra cuyo significado está incluido en el de otras.

3.3. Word embeddings[10]

Zellig Harris (1954) plantea la célebre idea de que *palabras con similar contexto poseen significados similares*. Esta idea se deriva de su pregunta de si el lenguaje posee una estructura de distribución, donde la distribución de un elemento se entenderá como la suma de todos sus entornos. Siendo el entorno de un elemento A el conjunto existente de sus co-ocurrentes, es decir, los otros elementos, cada uno en una posición particular, con la que A ocurre para producir un enunciado.

La hipótesis de que el lenguaje posee una estructura de distribución motiva la idea de encontrar o generar una estructura distribucional con la cual se pueda modelar el lenguaje, con lo cual nacen los modelos de *word embeddings*. Estos modelos son representaciones vectoriales que *capturan el significado de una palabra*, lo que sin duda es una idea difusa, pero que en la práctica se traduce en que son vectores en un espacio de alta dimensión de las palabras que componen el lenguaje, donde embeddings de palabras similares o relacionadas están cerca unas de otras y embeddings de diferentes palabras poseen una distancia mayor en este espacio vectorial.

La tarea siempre es factorizar una matriz de factores term-term que contenga recuentos de co-ocurrencia, información mutua puntual (PMI) o métricas similares. Las matrices de factores generalmente se llaman U y V, que definen dos dimensiones de embeddings distintos. U es una matriz que contiene los word embeddings finales y V es un conjunto temporal de word embeddings que contiene las representaciones utilizadas para las palabras. Esta operación (generalmente un proceso de optimización) se realiza escaneando el corpus[11] con una ventana de tamaño fijo. La ventana tiene una palabra central, llamada el objetivo y unas pocas palabras vecinas que se llaman contexto. Ambos: la palabra central y el contexto, se inicializan aleatoriamente en U y V, respectivamente. El objetivo es minimizar la distancia, medido por lo general mediante producto punto entre las palabras y sus contextos. Esto se hace realizando un descenso de gradiente estocástico, donde cada muestra estocástica es una ventana consecutiva en el corpus. Cada vez que se encuentra una palabra objetivo con un contexto, sus vectores se juntan ligeramente en proporción con la tasa de aprendizaje. Estos modelos son bastantes simples y, por lo tanto, muy eficientes. Se pueden ejecutar sobre corpus muy grandes y sólo requieren algunas iteraciones para lograr la convergencia.

3.3.1. Word2Vec

Word2Vec es probablemente el algoritmo de word embeddings más popular. No fue el primero, ya que existen modelos como el The Neural Network Language Model (NNLM)

[10]En esta sección, así como en la siguiente, mi trabajo esta fuertamente basado en la tesis de JE Alvarez (2017), quien rehalizó un excelente trabajo comparando algoritmos de similitud aplicado a textos académicos. Este trabajo posee una finalidad distinta, por lo que sus experimentos y resultados difieren de lo que se presentará en las siguientes secciones, pero su marco teórico y revisión de la literatura han sido un gran alivio en el desarrollo de este documento.

[11] La RAE define la palabra corpus como: "conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación".

(Bengio et al, 2003), pero si el primero de esta nueva generación de algoritmos de menor complejidad computacional que permiten reducir el tiempo de entrenamiento e incrementar el tamaño del corpus que se puede analizar.

El modelo planteado por Mikolov, K. Chen, et al. (2013) comienza con un conjunto de vectores de palabras que se inicializan aleatoriamente. Luego, escanea el corpus secuencialmente, manteniendo siempre una ventana de contexto alrededor de cada palabra objetivo. En este punto, hay algunas diferencias entre 2 modelos: el BoW[12] y Skip-gram[13], pero en esencia, el algoritmo calcula el producto de puntos entre la palabra objetivo y las palabras de contexto e intenta minimizar esta métrica realizando el Descenso del Gradiente Estocástico (SGD). Cada vez que se encuentran dos palabras en un contexto similar, se refuerza su vínculo o distancia espacial. Mientras más evidencia se encuentre al escanear el corpus de que dos palabras son similares, más cerca estarán.

El modelo básico que acabamos de describir solo proporciona un refuerzo positivo para acercar los vectores. Con un corpus infinito, el estado final sería que todos los vectores estarían en la misma posición, lo que obviamente no es el efecto deseado. Para abordar problema, inicialmente se propuso un regulador *Hierarchical Softmax*. Más tarde, propusieron un método alternativo llamado *Negative Sampling*. Este último es más simple y se ha demostrado que es más efectivo. La premisa básica es que cada vez que se reduce al mínimo la distancia entre vectores, se muestrean unas pocas palabras aleatorias y se maximiza su distancia al vector objetivo. De esta manera, se garantiza que las palabras no similares se mantengan lejos una de la otra.

Formalmente (y para el caso del BoW), Word2Vec considera un corpus, una secuencia de palabras $w_1, w_2, ..., w_T$ y un contexto de rango c (c palabras hacia la izquierda y hacia la derecha de la palabra objetivo). Así la función objetivo es:

$$max \ \frac{1}{T}\sum_{t=1}^{T} logP(w_t|\sum_{-c\leq j\leq c, j\neq 0} w_{t+j}) \tag{3.3}$$

La ecuación anterior se interpreta cómo la maximización de la probabilidad de ocurrencia de una palabra, dado su contexto. Por otro lado, la probabilidad de ocurrencia se cálcula como una Softmax, donde u_w es el word embedding para la palabra objetivo (w) y v_w es el word embedding del contexto. Lo anterior corresponde a la siguiende definición:

$$P(w_t|w_c) = \frac{\mathrm{e}xp(v_{w_c}^T u_{w_t})}{\sum_{w=1}^{W} \mathrm{e}xp(v_w^T u_{w_t})} \tag{3.4}$$

Sin embargo, Softmax es demasiado costoso como función de pérdida, ya que el cálculo

[12]BoW es el acrónimo de *Bag-of-Words*, es decir, bolsa de palabras o Term-Document Matrix 3.1.2

[13]No nos referiremos mayormente a esta variación del modelo, pero en esencia diremos que es otra forma de representar como vectores las palabras en un documento (así como lo es el Term-Document Matrix). En esencia, el Skip-Gram representa una palabra como pares (i, j) donde i es la palabra objetivo y j es una de las palabras de contexto. Así, para una ventana de 2 palabras, representar la palabra 'Había' en la oración 'Había una vez' implica la formación de los pares {(Había, una), (Había, vez)}.

del gradiente tiene una complejidad proporcional al tamaño del vocabulario W. El segundo documento (Mikolov, Sutskever, et al., 2013) propone dos soluciones. Una es usar Hierarchical Softmax, que es un algoritmo $O(log2W)$ para estimar Softmax. La segunda alternativa, y la más popular, es el Negative Sampling. La siguiente es la función objetivo por ventana que se maximizará en el caso de Negative Sampling.

$$log\ \ \sigma(v_{w_c}^T u_{w_t}) + \sum_{\mathrm{i}=1}^{k} \mathbb{E}_{w_\mathrm{i}} \equiv P_n(w)[log\ \ \sigma(-v_{w_c}^T u_{w_t})] \tag{3.5}$$

Donde k es un hiperparámetro que especifica el número de *negative samplings* aleatorias para usar en contraste con el modelo de aproximación entre los embeddings del objetivo y el contexto. Las *negative samplings* se extraen de la distribución $P_n(w)$.

3.3.2. GloVe

En el documento original (Pennington, Socher y C. Manning, 2014) los autores distinguen dos familias de modelos para la formación de word embeddings: métodos de factorización de matriz global (como LSA3.2.1) y métodos de 'ventana de contexto local' (como Word2Vec). Ellos afirman que ambos sufren importantes inconvenientes. Si bien los métodos de ventana de contexto como Word2Vec funcionan bien en la tarea de computo de similitud entre palabras o documentos, no aprovechan las estadísticas globales de co-ocurrencia de palabras. La principal motivación para GloVe es encontrar un término medio: un algoritmo que actúa sobre las estadísticas globales, pero logra la misma estructura semántica del espacio vectorial que Word2Vec. GloVe también muestra la razón por la cual se crean estos tipos de estructuras de vectores semánticos, haciendo explícitas las propiedades de las probabilidades de concurrencia. Así, su principal medida de similitud es la probabilidad de co-ocurrencia. Esto se puede entender mejor con un ejemplo, y el documento original brinda uno excelente.

Considere dos palabras relacionadas i = *hielo* y j = *vapor* (ver la tabla 3.1). Podemos examinar la relación entre estas palabras al observar su coincidencia con un conjunto de k palabras de prueba, $P_{\mathrm{i}k}/P_{jk}$. Si imaginamos quei y j definen un eje semántico de estado físico, podemos observar que la palabra k = *sólido* se observa que la relación es grande, lo que significa que está fuertemente posicionada en el extremo positivo de la escala. Del mismo modo, si observamos la palabra k = *gas*, donde la fracción pequeña, lo que indica que está semánticamente en el otro extremo de la escala. Las palabras no relacionadas, como k = *moda*, mostrarán un valor cercano a uno. Lo mismo ocurre con la palabra k = *agua*, que obviamente está muy relacionada con ambas palabras, pero es redundante e irrelevante expresar la relación entre ambas palabras, que es, de nuevo, el grado del estado físico de la materia.

GloVe formaliza el fenómeno descrito anteriormente y entrena los embeddings para que simulen dicha estructura. Esto puede parecer un poco intrincado al principio, pero en la formalización se puede ver claramente que la ecuación inicial conduce directamente a optimizar el producto escalar entre un vector de palabra y sus vectores de contexto, para que sea lo

Probabilidad y Ratio	$k = solido$	$k = gas$	$k = agua$	$k = moda$
$P(k\|hielo)$	$1{,}9x10^{-4}$	$6{,}6x10^{-5}$	$3{,}0x10^{-3}$	$1{,}7x10^{-5}$
$P(k\|gas)$	$2{,}2x10^{-5}$	$7{,}8x10^{-4}$	$2{,}2x10^{-3}$	$1{,}8x10^{-5}$
$P(k\|hielo)/P(k\|gas)$	8,9	$8{,}5x10^{-2}$	1,36	0,96

Tabla 3.1: Ejemplo del paper original Pennington, Socher y C. Manning (2014)

más cercano posible a su probabilidad de co-ocurrencia.

Arriba se ha descrito una propiedad de co-ocurrencia en probabilidades que puede usarse para definir relaciones semánticas. Hemos establecido un eje semántico con las palabras i = *hielo* y $j = vapor$ y las ponemos contra un conjunto de palabras conocidas k para observar cómo se relacionan con este eje. Este concepto se formaliza con la siguiente ecuación:

$$F(w_{\mathrm{i}}, w_j, \hat{w_k}) = \frac{P_{\mathrm{i}k}}{P_{jk}} \tag{3.6}$$

En la ecuación 3.1 una función F aún indefinida debe aplicarse a los vectores de palabras y la salida de F debe aproximarse a la fracción de probabilidad (que es una versión primitiva de la función objetivo que resuelve GloVe). Para esto, F se reduce a tomar un único parámetro escalar. Luego, dada la naturaleza lineal inherente de un espacio vectorial, las operaciones de diferencia y de producto de punto se eligen para combinar los argumentos.

$$F((w_{\mathrm{i}} - w_j)^T \hat{w_k}) = \frac{P_{\mathrm{i}k}}{P_{jk}} \tag{3.7}$$

Si se requiere que F sea un homomorfismo entre $(\mathbb{R}, +)$ y $(\mathbb{R}, X)$, F sólo puede ser la función $F = exp$. Luego de un par de simplificaciones, la ecuacion queda definida cómo:

$$w_{\mathrm{i}}^T \hat{w_k} = logP_{\mathrm{i}k} = logX_{\mathrm{i}k} - logX_{\mathrm{i}} \tag{3.8}$$

Donde X_{i} es la frecuencia total de la palabra i y $X_{\mathrm{i}k}$ es el recuento de las instancias en las que i se encuentra en el contexto de k. Esta ecuación se puede simplificar aún más al observar que $logX_{\mathrm{i}}$ no depende de k, por lo que se reemplaza por un vaor aleatorio (*bias*) b_{i}. Para la simetría, se agrega otro sesgo $\hat{b_k}$.

$$w_{\mathrm{i}}^T \hat{w_k} + b_{\mathrm{i}} + \hat{b_{\mathrm{i}}} = logX_{\mathrm{i}k} \tag{3.9}$$

Esta ecuación ya es sencilla de resolver, pero hay dos factores más que podrían ser mejorados. Primero, $logX_{\mathrm{i}k}$diverge cuando no hay co-ocurrencias entre i y k. Para solucionar esto, el término $logx$ es reemplazado por $log(1 + x)$. En segundo lugar, las co-ocurrencias muy frecuentes no son demasiado relevantes y en su mayoría contribuyen a aumentar el ruido. Por tanto, se agrega una función de ponderación $f(X_{\mathrm{i}j})$ para compensar esto. Hay muchas opciones para tal f pero los autores convergieron en la siguiente función.

$$f(x) = \begin{cases} (x/x_{max})^{\alpha}, & \text{si } x \leq x_{max} \\ 1, & \text{de lo contrario} \end{cases} \tag{3.10}$$

Los autores fijaron $x_{max} = 100$ y encontraron empíricamente el valor más apropiado para $\alpha = 3/4$. Curiosamente, se encontró que la misma escala era la mejor para Word2Vec. El último paso es simplemente convertir la ecuación en un problema de optimización de mínimos cuadrados. Este es el modelo final de GloVe.

$$J = \sum_{\mathrm{i},j=1}^{V} f(X_{\mathrm{i}j})(w_{\mathrm{i}}^{T}\hat{w_k} + b_{\mathrm{i}} + \hat{b_{\mathrm{i}}} - log(1 + X_{\mathrm{i}k}))^2 \tag{3.11}$$

3.3.3. FastText

Publicado a comienzos del 2017 por un grupo bajo la tutela del mismo autor que desarrolló Word2Vec, el algoritmo FastText (Bojanowski et al., 2016) es básicamente una extensión de Word2Vec que considera que cada palabra esta compuesta por n-grams (o subpalabras) que luego se combinan mediante una función de composición simple para calcular los embeddings de palabras finales. Dado este enfoque, FastText recibe la denominación de embedding modular.

El enfoque de este algoritmo implica que cada componente del vocabulario se repetirá más, por lo que se necesitan menos datos de entrenamiento. Otra gran ventaja es que la información de las subpalabras, como las variaciones morfológicas[14], se capturan correctamente. Otros algoritmos solo toman tokens estándar como palabras, y pueden crear diferentes embeddings para variaciones morfológicas, lo que aumenta el ruido. En el caso de FastText, las variaciones morfológicas mantienen la mayoría de sus componentes comunes y tienen ligeras modificaciones en sus embeddings basado en las diferencias como prefijos o sufijos. Usar n-grams para capturar características morfológicas puede parecer básico o poco sofisticado, ya que hay otros modelos que hacen una segmentación morfológica explícita, sin embargo, esto es intencional, ya que la simplicidad del método también aumenta la generalidad. Es gracias a esta generalidad que FastText funciona bien en idiomas muy diferentes cómo por ejemplo el español donde existen muchas variaciones morfológicas.

Formalmente, dado un gran cuerpo de entrenamiento representado como una secuencia de palabras $w_1, .., w_T$, el objetivo del modelo skipgram es maximizar la siguiente log-likelihood:

$$\sum_{t=1}^{T} \sum_{c \in C_t} logP(w_c|w_t) \tag{3.12}$$

Donde C_t es el contexto, es decir, el conjunto de palabras que se posicionan alrededor de

[14]La variación morfológica de una palabra hace referencia a las variaciones en el tiempo, género o singularidad/pluralidad a la que dicha palabra puede ser sometida.

la palabra w_t. Por otro lado, una posible opción para definir la probabilidad de una palabra de contexto es utilizar una Softmax (cómo en Word2Vec), no obstante, dicho modelo no se adapta a este caso, ya que dado una palabra w_t, la softmax sólo predice una palabra de contexto w_c. Por tanto, el problema de predecir palabras de contexto puede enmarcarse como un conjunto de tareas de clasificación binarias independientes. Así, el objetivo es predecir de forma independiente la presencia (o ausencia) de palabras contextuales. Para la palabra en la posición t consideramos todas las palabras de contexto como ejemplos positivos, y como ejemplos negativos muestras al azar del diccionario. Para una palabra de contexto en la posición c, utilizando la pérdida logística binaria, obtenemos la siguiente log-likelihood negativa:

$$log\Big(1 + \mathrm{e}^{-s(w_t,w_c)}\Big) + \sum_{n \in N_{t,c}} log\big(1 + \mathrm{e}^{s(w_t,n)}\big) \tag{3.13}$$

Donde $N_{t,c}$ es un conjunto de ejemplos negativos muestreados del vocabulario. Al denotar la función de pérdida logística como $\ell : x \rightarrow log(1 + \mathrm{e}^{-x})$, podemos reescribir la función objetivo como:

$$\sum_{t=1}^{T} \Big[\sum_{c \in C_t} \ell(s(w_t, w_c) + \sum_{n \in N_{t,c}} \ell(-s(w_t, n)) \Big] \tag{3.14}$$

Donde $s(w_t, w_c)$ es la función de similitud de los mismos vectores (embeddings) definidos en el modelo Word2Vec, usualmente llamados vectores de *input* y *output*. Por tanto, $s(w_t, w_c) = (u_{w_t}^T v_{w_c})$, al igual que en la versión de skip-gram del Word2Vec.

No obstante, al usar una representación vectorial distinta para cada palabra, el modelo de skip-gram ignora la estructura interna de las palabras. Por lo que en FastText se proponemos una función de similitud diferente para tener en cuenta esta información. Ahora, cada palabra w se representa como una bolsa de caracteres n-gram. En la notación original del paper, se añaden símbolos de límites especiales < y > al principio y al final de las palabras, lo que permite distinguir prefijos y sufijos de otras secuencias de caracteres. También se incluye la palabra w en el conjunto de sus n-gramas, para aprender una representación de cada palabra (además de los n-gramas de la palabra). En el paper original se muestra un ejemplo con la palabra *where* y $n = 3$, como ejemplo esta palabra será representada por los n-grams:

```
<wh, whe, her, ere, re>
```

y la secuencia especial

```
<where>
```

No obstante, los autores advierten que la secuencia `<her>`, que corresponde a la palabra *her* es diferente del tri-gram de la palabra *where*. En la práctica, se extraen todos los n-grams

Campo	Word2Vec	Glove	FasText
Considera el modelo la co-ocurrencia de palabras	No	Si	No
Uso de memoria entrenados sobre el SBWC[15]	708 mb	906 mb	802 mb
Accuracy para *analogy task*[16] en el estado del arte	89 %	87 %	93 %

Tabla 3.2: Tabla de comparación para Word Embeddings

para n mayor o igual a 3 y menor o igual a 6. Este es un enfoque muy simple, y se podrían considerar diferentes conjuntos de n-gramas, por ejemplo tomando todos los prefijos y sufijos.

Supongamos que se brinda un diccionario de n-grams de tamaño G. Dada una palabra w, denotemos por $\mathcal{G}_w \subset \{1, .., G\}$ el conjunto de n-grams que aparece en w. Luego, se asocia una representación vectorial z_g a cada n-gram g. Así, se representa una palabra por la suma de las representaciones de vectores de sus n-grams. Obtenemos así la función de similitud entre una palabra y su contexto como:

$$s(w, c) = \sum_{g \in \mathcal{G}_w} z_g^T v_c \tag{3.15}$$

Este simple modelo permite compartir las representaciones entre las palabras, lo que permite aprender una representación confiable de palabras extrañas.

La equación 3.15 formaliza la forma en que FastText se diferencia de Word2Vec al incoporar los n-grams de las palabras y posteriormente calcular una función de similitud entre las palabras y su contexto como el producto punto entre el embedding de los n-grams (z_g) y el embedding del contexto (v_c), que es exactamente lo que hace el modelo anterior pero con una granularidad menor.

3.3.4. Comparación Word Embeddings

Finalmente, en la Tabla 3.2 queremos resumir brevemente algunos aspectos relevantes para comparar estas 3 formas de representación vectorial para texto.

3.4. Modelos de similitud semántica textual

3.4.1. Doc2Vec

Este modelo es también conocido como Paragraph Vector (Quoc V. Le y Mikolov, 2014) y fue propuesto por el mismo equipo responsable de Word2Vec. Doc2Vec es considerado en este trabajo ya que fue uno de los primeros modelos capaces de construir representaciones de secuencias de entrada de longitud variable. A diferencia de algunos de los enfoques anteriores, es general y aplicable a textos de cualquier longitud: oraciones, párrafos y documentos.

En este modelo, cada párrafo[17] es asignado a un vector único, representado por una columna en una matriz D, y cada palabra también se mapea a un vector único, representado por una columna en la matriz W (al igual que Word2Vec). El vector de párrafo y los vectores de palabras se promedian o se concatenan para predecir la siguiente palabra en un contexto. En la Figura 3.7 se aprecia la extensión al modelo Word2Vec para mapear los párrafos a los que corresponden las palabras.

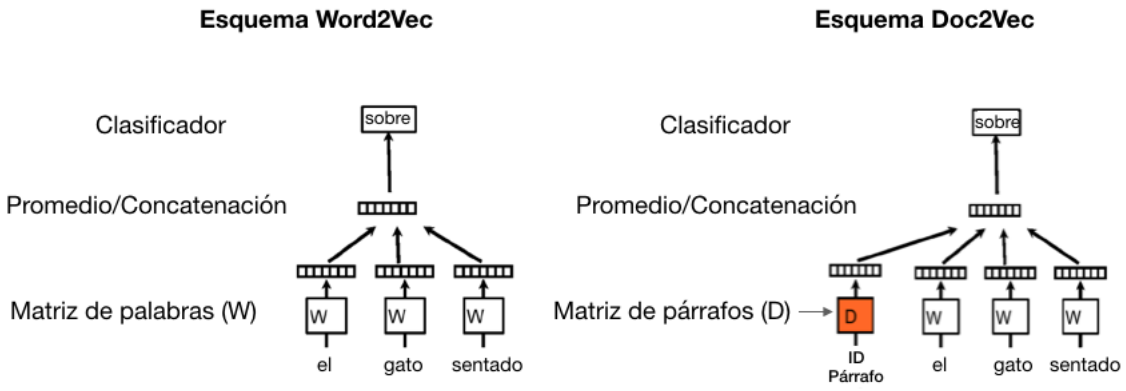


Figura 3.7: Representación de Doc2Vec y Word2Vec (Quoc V. Le y Mikolov, 2014)

En Doc2Vec, el token de cada párrafo se puede considerar como otra palabra que actúa como una memoria que recuerda lo que falta en el contexto actual. Por esta razón, los autores denominaron a este modelo de memoria distribuida de vectores de párrafo, o en inglés Distributed Memory Model of Paragraph Vectors (PV-DM).

Los contextos son de longitud fija y se muestrean con una ventana deslizante sobre los párrafos. El vector de los párrafos se comparte en todos los contextos generados a partir del mismo párrafo, pero no a través de los párrafos. Sin embargo, la matriz de palabras (W) se comparte entre párrafos, es decir, el embedding para 'poderoso' es el mismo para todos los párrafos.

Formalmente, los vectores de párrafos y palabras se estiman mediante el uso del descenso del gradiente estocástico y Backpropagation. Para esto, se maximiza la misma probabilidad que en Word2Vec (Ver ecuación 3.3), pero con la diferencia que en este caso se debe incorporar independencia de los párrafos. Así, dado un número M de documentos o párrafos, la función objetivo de este modelo es la que se aprecia en la ecuación 3.16.

$$max \ \frac{1}{M}\sum_{\mathrm{i}=1}^{M}\frac{1}{|D_{\mathrm{i}}|}\sum_{t=k}^{|D_{\mathrm{i}-1}|-k} logP(w_{\mathrm{i},t}|w_{\mathrm{i},t-k},..,w_{\mathrm{i},t+k};D_{\mathrm{i}}) \tag{3.16}$$

[17]En el paper original se utiliza el término párrafo para referirse a estos documentos de largo variable, pero en general se entenderá por párrafo un documento o composición de párrafos.

3.4.2. Word Mover's Distance

Este método es probablemente el que mayor afinidad posee con el área de investigación operaciones y con la carrera de Ingeniería Industrial, ya que resuelve un problema de transporte mediante programación lineal. El objetivo es transformar (mover) de forma óptima las palabras de un documento para que se conviertan en un segundo documento. Si dos documentos son semánticamente diferentes, sus embeddings estarán lejos uno de otro, por lo que el costo será alto. Lo contrario sucede si los documentos son similares.

La implementación original de este modelo fue basado en Word2Vec, en donde aprovechando la representación vectorial que ofrece este modelo, computa la distancia entre los embeddings de 2 documentos. Una de esas medidas de disimilitud entre palabras es naturalmente proporcionada por su distancia euclidiana en el espacio vectorial de sus embeddings. Más precisamente, la distancia entre la palabra i y la palabra j será denotado cómo $c(\mathrm{i}, j) = \|x_{\mathrm{i}} - x_j\|_2$. Donde $x_{\mathrm{i}} \in \mathbb{R}^{\mathrm{d}}$, y $\mathrm{d} \in \mathbb{R}^n$ es la representación vectorial de los documentos asociada a la Term-Document Matrix 3.1.2. Para ser precisos, si una palabra i aparece k_{i} veces en un documento, se denota $\mathrm{d}_{\mathrm{i}} = \frac{n_{\mathrm{i}}}{\sum_{j=1}^{n} k_j}$.

Así, el 'costo de transporte' entre dos palabras es la unidad básica para crear una distancia entre dos documentos. Definamos así como d y d′ la representación vectorial de dos documentos en término de sus vectores en la Term-Document Matrix. Primero, se permite que cada palabra i en d se transforme en cualquier palabra en d′[18]. Definamos $T \in \mathbb{R}^{nxn}$ cómo la matriz de flujo (sparse) donde $T_{\mathrm{i}j} \geqslant 0$ denota cuánto de la palabra i en d viaja a la palabra j en d′. Para transformar d completamente en d′, es necesario asegurar que todo el flujo de salida de la palabra i es igual a d_{i}, es decir, $\sum_j T_{\mathrm{i}j} = \mathrm{d}_{\mathrm{i}}$. Además, la cantidad de flujo entrante a la palabra j debe coincidir con d'_j, es decir, $\sum_j T_{\mathrm{i}j} = \mathrm{d}'_{\mathrm{i}}$. Finalmente, podemos definir la distancia entre los dos documentos como el costo acumulado mínimo (ponderado) requerido para mover todas las palabras de d a d′, es decir, $\sum_{\mathrm{i},j} T_{\mathrm{i}j} c(\mathrm{i}, j)$.

Formalmente, el costo acumulado mínimo de mover d a d′ dadas sus restricciones, lo proporciona la solución al problema lineal de la ecuación 3.17. Este problema de optimización es un caso especial de la métrica *earth mover's distance (EMD)* (Monge, 1781; Nemhauser y Wolsey, 1988; Rubner et al, 1998).

$$\begin{aligned} & \min_{T \geqslant 0} \sum_{\mathrm{i},j=1}^{n} T_{\mathrm{i},j} c(\mathrm{i}, j) \\ \text{Sujeto a} : & \sum_{j=1}^{n} T_{\mathrm{i}j} = \mathrm{d}_{\mathrm{i}}, \forall \mathrm{i} \in \{1, .., n\} \\ & \sum_{\mathrm{i}=1}^{n} T_{\mathrm{i}j} = \mathrm{d}'_j, \forall j \in \{1, .., n\} \end{aligned} \tag{3.17}$$

La solución óptima es que cada palabra en d mueva toda su masa de probabilidad a la

[18]Notar que dado esta condición, WMD no tiene en cuenta el orden de las palabras, por lo que codifica ambos documentos según la Term-Document Matrix.

Corpus	WMD	Doc2Vec
Títulos	90 %	65 %
Abstracts	92 %	86 %
Cuerpo	NA[19]	97 %

Tabla 3.3: Tabla de comparación para algoritmos de similitud semántica textual (JE Alvarez, 2017)

palabra más similar en d′. Así, una matriz T^* óptima se define cómo:

$$T^*_{\mathrm{i},j} = \begin{cases} \mathrm{d_i}, & \text{si } j = argmin_j c(\mathrm{i}, j) \\ 0, & \text{de lo contrario} \end{cases} \tag{3.18}$$

3.4.3. WMD vs Doc2vec

JE Alvarez (2017) analizó el desempeño de estos 2 algoritmos en conjunto con otros más en una labor muy similar: intentar agrupar documentos científicos en base a su área (por ejemplo, biología, astronomía, etc). El autor los comparó el desempeño para 3 distintos corpus: los título de papers, los absracts y el finalmente el cuerpo completo del documento. Los resultados son lo que se aprecian en la Tabla 3.3

Es necesario destacar que se escogió implementar estos 2 algoritmos en este trabajo precisamente por los resultados reportados en la Tabla 3.3, ya que de los 4 algoritms evaluados por JE Alvarez (2017), WMD fue el que obtuvo mejor desempeño en el corpus de abstracts y Doc2Vec el que obtuvo mejor desempeño para cuerpos de documentos. Nuestros relatos tienen una extensión promedio cercana a los 200 términos (post-preprocesamiento), por lo que la extensión es mayor a la de un abstract y bastante menor al cuerpo de un paper científico.

3.5. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA), (Blei et. al, 2003) es un modelo probabilístico generativo para colecciones de datos discretos tales como corpus de texto. LDA es un modelo bayesiano jerárquico de tres niveles, en el que cada elemento de una colección se modela como una mezcla finita sobre un conjunto subyacente de tópicos. Cada tópico, a su vez, se modela como una mezcla infinita sobre un conjunto subyacente de probabilidades de tópicos. En el contexto del modelado de texto, las probabilidades de los tópicos proporcionan una representación explícita de un documento (es decir, cada documento posee una probabilidad explícita de pertenecer a uno o más tópicos).

El objetivo de este modelo, y en general de los modelos de topicalización es encontrar descripciones breves de los miembros de una colección que permitan el procesamiento eficiente de grandes colecciones al mismo tiempo que se preservan las relaciones estadísticas esenciales

que son útiles para tareas básicas tales como clasificación, detección de novedades, resumen y juicios de similitud y relevancia.

Este modelo considera las siguientes definiciones:

- Una *palabra* es la unidad básica de los datos discretos, definido como un elemento de un vocabulario indexado por $\{1, ..., V\}$.
- Un *documento* es una secuencia de N palabras, denotado por: $w = (w_1, w_2, .., w_N)$.
- Un *corpus* es una secuencia de documentos denotado por: $D = \{w_1, w_2, .., w_M\}$.

LDA asume el siguiente proceso generativo para cada documento w en un corpus D:

1. $N \sim Poisson(\xi)$
2. $\theta \sim Dir(\alpha)$
3. Para cada una de las N palabras w_n
   - Escoger un tópico $z_n \sim Multinomial(\theta)$
   - Escoger una palabra w_n desde $P(w_n|z_n, \beta)$, una probabilidad multinomial condicionada al tópico z_n.

Dado los parámetros α y β, la distribución conjunta de una mezcla de tópicos θ, un conjunto de N tópcios z_n, y un conjunto de N palabras w, esta dada por:

$$p(\theta, z, w|\alpha, \beta) = p(\theta, \alpha) \prod_{\mathrm{i}=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.19}$$

3.6. Métricas de desempeño

3.6.1. Similitud Coseno

Dada una representación vectorial de los documentos y sus términos como la Term-Document Matrix 3.1.2, la similitud coseno puede ser aplicada para medir la similitud entre documentos dada la representación vectorial de estos. Esta es probablemente una de las primeras medidas de similitud que funcionaron relativamente bien, y cómo ha sido descrito anteriormente, esta la forma más convencional de recuperar información utilizando LSA 3.2.1.

La definición formal de la similitud coseno es que dado dos vectores de dimensión N, digamos $\overrightarrow{v}$ y $\overrightarrow{w}$ esta medida se calcula cómo:

$$Cosine(\overrightarrow{v}, \overrightarrow{w}) = \frac{\overrightarrow{v} \cdot \overrightarrow{w}}{|\overrightarrow{v}||\overrightarrow{w}|} = \frac{\sum_{\mathrm{i}=1}^{N} v_\mathrm{i} \times w_\mathrm{i}}{\sqrt{\sum_{\mathrm{i}=1}^{N} v_\mathrm{i}^2} \sqrt{\sum_{\mathrm{i}=1}^{N} w_\mathrm{i}^2}} \tag{3.20}$$

Esta medida fue extensamente utilizada en los inicios de la teoría de recuperación de la información, principalmente a través de una versión ponderada del Term-Document Ma-

		Valor predicho		
		p	**n**	**total**
Valor real	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Tabla 3.4: Matriz de confusión

trix, que fue previo al LSA. Esta versión fue llamada Term Frequency - Inverse Document Frequency (TF-IDF), y básicamente consiste en considerar el Term Frecuency, es decir, la cantidad de veces que aparece un término t aparece en un documento d, el cual se pondera por el Inverse Document Frequency, qué es un normalizador de la frecuencia de una palabra, lo cual provee una medida de cuán extraño es un término en una colección de documentos (N).

$$
\begin{aligned}
tf(t,\mathrm{d}) &= \begin{cases} 1, & \text{si} t \text{ aparece en d} \\ 0, & \text{de lo contrario} \end{cases} \\
\mathrm{id}f(t,D) &= log\frac{N}{|\{\mathrm{d} \in D : t \in \mathrm{d}\}|} \\
\text{TF-IDF}(t,\mathrm{d},D) &= tf(t,\mathrm{d}) \times \mathrm{id}f(t,D)
\end{aligned}
\tag{3.21}
$$

3.6.2. Precision

Concepto definido por Perry, Kent y Berry (1955) y en sus orígenes ampliamente utilizado por los motores de búsqueda en la web. La idea de esta medida es cuantificar el desempeño de un buscador (ya sea en términos de búsqueda de páginas web, documentos, etc.) en cuanto a la calidad de la información que recupera. El objetivo de un buscador es brindar información relevante que responda a una query. No obstante el buscador puede identificar de forma errónea la relevancia de la información y es de ese punto del que se hace cargo esta métrica.

Dada la Matriz de Confusión que se aprecia en la Tabla 3.4, la Precision se computa como:

$$
Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3.22}
$$

Así, la Precision se interpretará como la probabilidad de que un elemente haya sido correctamente recuperado, dado que el algoritmo lo recuperó.

3.6.3. Recall

Otro concepto introducido por Perry, Kent y Berry (1955) y cuya finalidad es similar a la de Precision. Recall intenta cuantificar el porcentage de resultados que fue correctamente recuperado, es decir, y dado nuevamente la Matriz de Confusión en la Tabla 3.4:

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3.23}$$

Así, Recall dará cuenta del porcentaje de elementos correctamente recuperados por el algoritmo.

3.6.4. F1-Measure

F1-Measure es el promedio armónico de Precision y Recall, donde F1-Measure alcanza su máximo valor en 1 y peor en 0. Usualmente acompaña las métricas de Precision y Recall, ya que estás métricas por si sólas sólo entregan una visión parcial de la recuperación de información. Así, en caso de que exista un *trade-off* entre Precision y Recall, será útil considerar el F1-Measure para balancear dichas asimetrías. El F1-Measure se computa cómo:

$$F1 - Measure = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.24}$$

3.6.5. Perplexity

Desarrollado por Jelinek et al. (1977), es una medida que –y explicado en su forma más general – permite cuantificar cuán bien una distribución de probabilidad predice una muestra. En el caso de subconjuntos de documentos o tópicos, Perplexity cuantifica cuanto se equivoca un modelo en elegir una palabra de la distribución general de palabras basado la entropía de la distribución subyacente. Por tanto, el objetivo es minimizar la más posible esta medida ya que indica que la distribución de probabilidad es buena para predecir la muestra (o tópicos).

Dado un corpus D de documentos, con T cómo el número de documentos (es decir, $n(D) = T$) y N el número de palabras (w) en el corpus, Perplexity se computa como se muestra en la siguiente ecuación:

$$Perplexity(D) = \mathrm{e}xp\Big(- \frac{\sum_{\mathrm{d}=1}^{T} logP(w_{\mathrm{d}})}{\sum_{\mathrm{d}=1}^{T} N_{\mathrm{d}}} \Big) \tag{3.25}$$

3.6.6. Topic Coherence

Perplexity es un esfuerzo por intentar definir una métrica de desempeño sobre la capacidad de predecir un tópico a partir de una distribución de probabilidad, no obstante, esta métrica no se correlaciona directamente con el juicio e interpretación de los humanos. Así, surge la motivación de buscar nuevas métricas que intenten replicar el juicio humano. No obstante, esta es una tarea difícil ya que el juicio humano no está claramente definido; por ejemplo, dos expertos pueden estar en desacuerdo sobre la utilidad de un tópico o tema.

La idea detrás de esta métrica desarrollada por Mimno et al. (2011), es que palabras que pertenecen a un mismo concepto van a co-ocurrir de forma frecuente en los documentos. Así, y utilizando conocimiento experto de humanos definieron y validaron tres niveles de tópicos: *buenos*, *intermedios* y *malos*. Dada esta definición, lo esperado es que en tópicos *buenos* e *intermedios* pares de palabras que pertenecen a un solo concepto coincidan en un mismo documento (por ejemplo, 'nucleicos' y 'ácidos' en documentos sobre el ADN), y pares de palabras que pertenecen a diferentes conceptos (por ejemplo, 'graso' y 'nucleico') no lo harán. Así, dada $D(v)$ la frecuencia del documento que contiene la palabra v (es decir, el número de documentos con al menos un token de tipo v) y $D(v, v')$, la cantidad de documentos donde co-ocurren los términos v y v' (es decir, el número de documentos que contienen uno o más tokens de tipo v y al menos un token de tipo v'), se define Topic Coherence de un tópico cómo:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{3.26}$$

Donde, $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ es la lista de las M palabras más frecuentes en el tópico t.

Capítulo 4

Desarrollo metodológico

En este capítulo se describe el actual proceso de persecución penal con el objetivo de entender el contexto del proyecto y el impacto que podría generar. Luego se describe los datos que ha sido brindada el Ministerio Público para este trabajo, su procesamiento y características. Posteriormente se señala cómo son implementados los algoritmos de recuperación de documentos, similitud semántica textual y topicalización.

4.1. Entendimiento del contexto

De acuerdo a la señalado en los artículos 79 y 80 del Código Procesal Penal, Carabineros de Chile y Policía de Investigaciones son los órganos auxiliares o colaboradores del Ministerio Público en las tareas de investigación criminal, realización de diligencias cuando así se decretare, bajo la dirección y responsabilidad del Ministerio Público. Por tanto, la denuncia de un delito puede ser realizada a través de estas instituciones, así como los tribunales con competencia criminal.

Como se aprecia en la Figura 4.1, una vez recibida una denuncia, se asigna un fiscal, que son abogados que forman parte del personal de la fiscalía y, de acuerdo a la ley, son los encargados de dirigir la investigación de los hechos constitutivos de delito, ejercer la acción penal pública, y proteger a las víctimas y testigos en todas aquellas causas que se les asigna. La fiscalía que investigará la causa será aquella que tenga competencia en la comuna donde presuntamente fue cometido el delito. Además, la fiscalía asigna la denuncia a un fiscal de acuerdo a su especialidad (delitos sexuales, económicos, drogas, VIF, etc.) o de acuerdo a la complejidad que tenga la investigación.

Una vez asignado un fiscal a la causa, se inicia la investigación, en la que el fiscal revisa los antecedentes, y si estos pueden ser constitutivos de un delito, abre una investigación en la que podrán ser citados a declarar tanto la víctima como los testigos a quienes se les consultará sobre los hechos que se investigan. Además, puede decretar la práctica de órdenes de investigar o instrucciones particulares a las policías, y la realización de peritajes por organismos especializados. Si el fiscal durante la investigación reúne antecedentes que

Figura 4.1: Proceso de persecución penal

le permitan llevar a juicio al o los presuntos responsables, formulará acusación – previa formalización - contra ellos para luego presentar las pruebas en un juicio oral y público, donde un tribunal competente decidirá si condena o absuelve a los imputados.

El cierre de una investigación significa que el fiscal ha terminado su labor investigativa referida al establecimiento del hecho delictual y a la participación del imputado en éste. El fiscal tiene un plazo para desarrollar la investigación que se cuenta desde la formalización. El plazo máximo que establece la ley es de dos años, y el juez de garantía puede resolver otorgar un plazo inferior. Finalizado el plazo, el fiscal debe cerrar la investigación y puede formular acusación en contra del imputado, o bien, proponer al juez el sobreseimiento temporal o definitivo, o **hacer uso de la facultad de no perseverar, si no cuenta con antecedentes suficientes para formular acusación**. En la sección 4.1.1 se muestran algunas estadísticas para dimensionar en que porcentaje de las investigaciones, no perseverar parece ser la única solución si es que no existen mayores antecedentes; en particular, la cifra que da cuenta del número de veces que un delito cuenta con un autor (o imputado) desconocido es bastante preocupante.

Finalmente, las causas pueden terminar de manera jurisdiccional (en tribunales) o facultativa (por decisión interna de la Fiscalía). Jurisdiccionalmente existen los siguientes tipos de términos: sentencia, suspensión condicional, acuerdo reparatorio, sobreseimiento definitivo o temporal, archivo provisional[1], principio de oportunidad y facultad de no inicio.

[1]Archivo Provisional significa que el caso es archivado, provisionalmente por no contar con antecedentes que permitan desarrollar una investigación para saber cómo ocurrieron los hechos y quiénes fueron los culpables. Sin embargo, se podrá reactivar o reabrir la investigación en cuanto surjan nuevos antecedentes.

4.1.1. Estadísticas

Desde el año 2015, la Fiscalía ha publicado las estadísticas más relevantes acerca de los delitos que fueron tramitados durante el año. De estos documentos se desprende valiosa información que da cuentas del nivel de resolución de los procesos de investigación llevados a cabo. Las cifras presentadas en esta sección corresponden a las reportadas por la institución en sus boletines estadísticos anuales disponibles en su página web[2].

De las Figuras **??** y 4.3 se puede apreciar que en los delitos contra la propiedad, la mayor proporción de estos son cometidos por imputados desconocidos, destacando los robos no violentos con un $92,20\,\%$ de los casos. Cabe destacar que en el año 2017, 404.471 causas de un total de 1.323.324, es decir, el $30,6\,\%$ fueron delitos de las categorías Robos, Robos no violentos y Otros delitos contra la propiedad, donde el $79,1\,\%$ de las 404.47 poseían un imputado desconocido.

A su vez, se puede apreciar en la Figura 4.2 que la cantidad de causas que terminan como un archivo provisional suman 688.335 de un total de 1.537.706 causas a las que se le aplico un término a su investigación, lo que representan el 44,76 % de los casos. Esta cifra es particularmente preocupante, ya que significa que en el año 2017 un total de 688.335 causas fueron archivadas provisionalmente por no contar con antecedentes que permitan desarrollar una investigación para saber cómo ocurrieron los hechos y quiénes fueron los culpables. Al respecto, la vocera de la Fiscalía Nacional, Marta Herrera señaló el año 2016 al diario La Tecera[3] que: "el tema del archivo está muy relacionado con el tipo de ingresos (de ilícitos) que tienen las regiones, es decir, las que tienen un mayor ingreso de delitos contra la propiedad deberían tener una mayor tasa de archivo, porque en esos ilícitos hemos encontrado más recurrentemente situaciones de imputado desconocido, sin poder dar con su identificación en los primeros momentos de la investigación, siendo más probable una alta tasa de archivos".

4.2. Comprensión los datos

Tal como se explicó en la sección anterior, el proceso de persecución penal comienza cuando una persona realiza una denuncia, cuyo relato del hecho es registrado manualmente por un funcionario según lo que el denunciante relata. Esto es almacenado en una base de datos que al ser migrado a la plataforma SIMAC del Ministerio Público cuenta con los campos que se aprecian en la Figura 4.1.

Cómo un mecanismo para priorizar la investigación de aquellos delitos contra la propiedad que generan mayor nivel de afectación a la población, el Sistema de Análisis Criminal y Focos Investigativos, creado por la Ley de Fortalecimiento del Ministerio Público, canaliza y agrupa causas cuya investigación adquiere una prioridad superior. En este caso, la Fiscalía Local de

[2]Los boletines pueden ser descargados desde el siguiente link: http://www.fiscaliadechile.cl/Fiscalia/estadisticas/index.do/

[3]Es posible acceder a la noticia completa titulada *Las razones del Ministerio Público para archivar cerca de la mitad de sus casos* a través del siguiente link: http://www2.latercera.com/noticia/las-razones-del-ministerio-publico-archivar-cerca-la-mitad-casos/. Información recuperada en Abril del 2018.

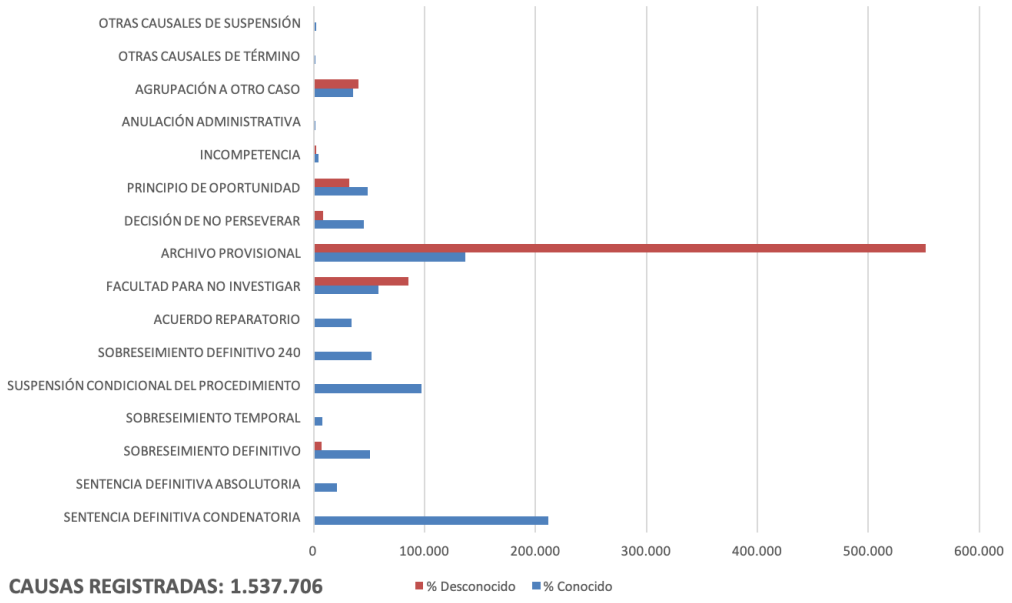


Figura 4.2: Número de delitos ingresados por categoría de delitos y tipo de imputado, año 2017

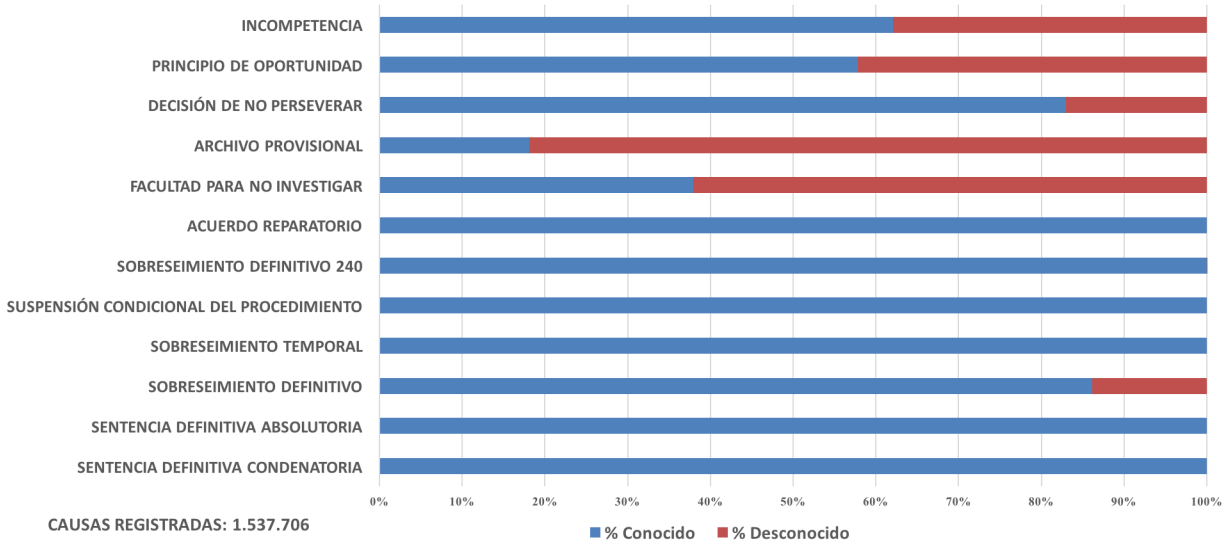


Figura 4.3: Términos para la investigación aplicados por tipo de imputado, año 2017

San Antonio ha declarado el foco de investigación Nº 23, que consta de 16 delitos[4] y crímenes cometidos entre el 24 de Febrero del 2018 y el 3 de Marzo del 2018.

[4]Los datos del foco de investigación Nº23 fueron brindados el día 19 de Abril del 2018, a ese día, el foco contaba de 16 delitos. Sin embargo, a través del seguimiento del foco es probable que la cantidad de registros de este foco haya aumentado.

Campo	Descripción
RUC	Rol único de la causa (ID)
Fecha Hecho	Fecha en que la víctima señala que fue cometido el delito
Fecha Recepción	Fecha en que se recibe la denuncia en alguna institución competente
Delito	Descripción del delito cometido
Lugar Ocurrencia	Lugar en que se comete el delito. Por ejemplo: Vía pública, Domicilio, etc.
Región	Región en que fue cometido el delito
Comuna	Comuna en la que se cometio el delito
Dirección del delito	Dirección de referencia en la que fue cometido el delito
Tipo Arma	Tipo de arma que se utilizó para cometer el delito (en caso de existir)
Relato	Descripción del delito que presta quién denuncia

Tabla 4.1: Campos de una causa ingresa a SIMAC

4.2.1. Registro de robos en lugar habitado Macrozona San Antonio

En la Macrozona de San Antonio se registraron 3.803 robos en lugar habitado o destinado a la habitación entre el 8 Enero del 2018 hasta 1 Abril del 2018. En la Tabla 4.5 se puede apreciar un ejemplo de los relatos que se encuentran en estos registros.

4.2.2. Foco de investigación Fiscalía Local San Antonio

El Foco Nº 23 establecido en la Macrozona de San Antonio contempla delitos de robo en lugar habitado y robos con violencia e intimidación en el lugar de mora. Se han asignado 16 causas a este foco y su las razones que vinculan a estos delitos en conjunto se desconocen, es por esto que con el objetivo de comprender este foco se han analizado de forma cualitativa.

En la Tabla 4.2 se describen los campos que uno o varios analistas completaron en la plataforma SIMAC.

Análisis cualitativo de las causas en el foco de investigación

Dado que las razones que por las cuales se ha constituido este foco son desconocidas, se procedió a leer y analizar cada una de estas con el objetivo de establecer similitudes y diferencias. En la Tabla 4.3 se puede apreciar una caracterización de los relatos basado en 6 atributos, estos son: si el delito se cometió en un inmueble habitado, el número de perpetradores, la presencia o ausencia de violencia en el delito, el uso de armas por parte de

Campo	Descripción
RUC	Rol único de la causa (ID)
FOCO	Nombre y número del foco de investigación declarado
DESCRIPCIÓN	Descripción extensa del nombre del foco
MACROZONA	Macrozona geográfica definida por Fiscalía
FISCALIA	Nombre de la Fiscalía donde se realiza la investigación
COMUNA	Comuna en la que se cometió el delito
FECHA_DELITO	Fecha en que se cometio el delito
FECHA_RECEPCION	Fecha de recepción de la causa
DELITO	Tipo de delito cometido
SITIO_SUCESO	Lugar de ocurrencia del delito. Por ejemplo: Vivienda, Tienda comercial, etc.
LUGAR_OCURRENCIA	Categoría para el sitio del suceso. Por ejemplo: Lugar habitado, Bien nacional público, etc.
GLOSA	Información relevante extraída de la glosa. Por ejemplo: existencia de cámara de vigilancia.
DATOS GENERALES DE LA CAUSA	Información parametrizada por el analista, se aprecia redundancia con campos anteriormente descritos
FORMAS DE COMISIÓN	Se registra información detallada del Modus Operandi
ARMADO(S)	Se registra información sobre las armas descritas en caso de existir
MEDIO DE TRANSPORTE	Medio de transporte de los delincuentes
MODALIDADES	Señala si hay información sobre un individuo o un grupo de estos
ESPECIE(S) SUSTRAIDA(S)	Especies sustraídas en el delito
AVALÚO SUSTRAÍDO	Avalúo de las especies sustraídas en el delito
MEDIO(S) DE PRUEBA - CÁMARAO	En caso de existir una cámara que haya registrado el suceso se detalla su ubicación y status sobre extracción de información
MEDIO(S) DE PRUEBA - EVIDENCIA	Se señala si hay evidencia extra cómo huellas dactilares
CANT_SOSPECHOSOS	Cantidad de sospechosos
DESC_SOSPECHOSOS	Descripción de los sospechosos
DATOS_SOSPECHOSO	Otros datos de los sospechosos
APODO_SOSPECHOSO	Apodo del sospechoso
TESTIGOS	En caso de existir testigos, su información se recopila en los siguientes 3 campos
DESC_TESTIGOS	Se caracteriza al testigo. Por ejemplo: llamada anónima
DATOS_TESTIGO	Información que aportan los testigos

Tabla 4.2: Campos relevantes que se encuentran en el foco de investigación

Tipo de delito	Nº de registros	Porcentaje
Robo en lugar destinado a la habitación	10	63 %
Robo en lugar habitado	6	37 %
Número de autores	Nº de registros	Porcentaje
Un individuo	10	63 %
Más de un individuo	6	37 %
Violencia e intimidación	Nº de registros	Porcentaje
Robo con violencia	3	19 %
Robo sin violencia	13	81 %
Presencia de armas	Nº de registros	Porcentaje
Robo con armas	3	19 %
Robo sin armas	13	81 %
Tipo de imputado	Nº de registros	Porcentaje
Conocido	4	25 %
Desconocido	12	75 %
Captura	Nº de registros	Porcentaje
Capturado	8	50 %
No capturado	8	50 %

Tabla 4.3: Análisis cualitativo de las causas en el foco de investigación

los delincuentes, si del relato se deduce un autor conocido y finalmente, si es que el autor del delito fue capturado tras el suceso o fue sorprendido de forma flagrante. En la Tabla 4.4 se detallan las especies que se declaran fueron sustraídas por el (los) delincuente(s).

4.2.3. Discusión

De los resultados expuestos en la Tabla 4.3 no se aprecian *modus operandis* particularmente similares o descripción de delincuentes que permitan sospechar de un autor común. Más aún, la categoría de autor conocido o desconocido es bastante similar a la distribución de todas las causas de delitos contra la propiedad del año 2017 (Ver Figura **??**. Probablemente, aquel atributo que destaca en esta muestra es la cantidad de causas en la que el autor fue capturado de forma inmediata o sorprendido de forma flagrante. En las especies sustraídas tampoco se aprecia un patrón distintivo, ya que los televisores son bien conocidos por ser un *hot product* (Clarke y Webb, 1999), por tanto, esta muestra sólo replica resultados conocidos.

Más allá de lo anterior, se aprecia una muestra de crímenes bastante heterogenea en sus formas de comisión, resultado y/o desenlace, lo cual no permite establecer una categoría o atributo predominante para el foco.

Especie sustraída	Frecuencia
Cilindro de gas	2
Videojuego	1
Ropa	1
Televisor	7
Herramientas	2
Camara fotográfica	1
Dinero en efectivo	2
Computador	1
Otras	5

Tabla 4.4: Especies sustraídad en las causas en el foco de investigación

4.3. Preparación de los datos

Tal como se describió en el capítulo anterior, una parte importante de trabajar con texto es su preprocesamiento. Gracias a la librería `NLTK` y las reglas de expresión este proceso es sumamente sencillo y rápido. En este caso, las acciones que se realizaron fueron:

- Remoción de stop words, números y caracteres especiales: este proceso fue el único que requirió retroalimentación de etapas posteriores del trabajo, ya que en la medida en que se analizaron los resultados se sumaban nuevas palabras como stop word. Se reemplazaron todas las letras con tilde por la misma, pero sin este y se eliminaron números y caracteres especiales.
- Transformación a minúscula y tokenización: una vez eliminados caracteres y stop words, estas fueron transformadas a minúsculas y se tokenizaron con el objetivo de poder trabajar con las palabras restantes con algoritmos.
- Stemming y Lematización: dado que en este trabajo se prioriza el uso de word embeddings, la lematización o stemming del corpus no es necesario, ya que los embeddings son capaces de representar con vectores similares a palabras similares en su contexto.

En la Tabla 4.5 se aprecia un ejemplo de un relato antes y después del preprocesamiento del texto.

4.4. Modelamiento

El proceso general que sigue este trabajo es el que se muestra en la Figura 4.4. El input del proceso son las denuncias en la forma de una colección de documentos (o corpus), le sigue su procesamiento, lo cuál fue descrito recientemente, luego de lo cuál sigue la aplicación de los modelos de recuperación de documentos por query y la aplicación de los algoritmos de similitud semántica textual (SST). De estos dos últimos puntos es precisamente de lo que se tratan los siguientes apartados.

Relato original	Relato post preprocesamiento
RELACION DE LOS HECHOS VIVO EN UNA HABITACION EN LA PROPIEDAD TIPO CITE, UBICADA EN CALLE TERESA N° 978, COMUNA DE LA CALERA. EL CASO ES QUE EL DIA DE AYER, A ESO DE LAS 11 : 00 HORAS, SALE DE MI CASA, REGRESANDO A ESO DE LAS 21 : 00 HORAS, PERCATANDOME QUE PERSONAS DESCONOCIDAS, FORZANDO LA VENTANA QUE DA AL PATIO DE LA CASA, HABTAN INGRESADO A MI HABITACION SUSTRAYENDO MI PENSION LA CUAL MANTENIA ENCIMA DEL VELADOR EN UNA CAJA, CORRESPONDIENTE A LA SUMA DE 150,000 PESOS (CIENTO CINCUENTA MIL PESOS), ESCAPANDO POR LA MISMA VENTANA EN DIRECCION DESCONOCIDA. QUISIERA AGREGAR QUE TRAS CONSULTAR A LOS VECINOS DEL CITE, DON JUAN TAPIA, DESCONOZCO MAYORES ANTECEDENTES ME SENAL6 QUE EL ESTABA DURMIENDO CUANDO ESCUCHO MUCHO RUIDO EN EL PATIO, SE ASOM6 Y VIO A CONSTANZA HERRERA CARVAJAL, SALIENDO POR LA VENTANA DE MI PIEZA. FINALMENTE QUISIERA INDICAR QUE CONSTANZA ES HIJA DE LA PAREJA DE DON MIGUEL CABRERA, TAMBIEN VECINO DEL CITE, DESCONOZCO MAYORES ANTECEDENTES, POR LO QUE TODOS EN LA PROPIEDAD LA CONOCEMOS, PERSONALMENTE SENALAR QUE EN VARIAS OPORTUNIDADES CONSTANZA HA IDO A MI HABITACION A SOLICITAR, DIVERSAS COSAS COMO LIMONES, TE, ETC., PERO JAMAS LE HE DADO LA CONFIANZA PARA QUE INGRESE ASI A MI PIEZA. POR ULTIMO INDICAR QUE SEGUN LO QUE AVERIGUE CONSTANZA ESTA VIVIENDO EN CALLE COCHRANE CON ACONCAGUA, POBLACION EL TRIGAL, COMUNA DE LA CALERA, AL COSTADO DE LA NUMERACION 878. ES TODD CUANTO PUEDO SENALAR AL RESPECTO	'vivo', 'habitacion', 'propiedad', 'tipo', 'cite', 'ubicada', 'teresa', 'calera', 'caso', 'ayer', 'sale', 'casa', 'regresando', 'percatandome', 'personas', 'desconocidas', 'forzando', 'ventana', 'da', 'patio', 'casa', 'habtan', 'ingresado', 'habitacion', 'sustrayendo', 'pension', 'mantenia', 'encima', 'velador', 'caja', 'correspondiente', 'suma', 'pesos', 'ciento', 'cincuenta', 'pesos', 'escapando', 'misma', 'ventana', 'direccion', 'desconocida', 'quisiera', 'agregar', 'tras', 'consultar', 'vecinos', 'cite', 'don', 'juan', 'tapia', 'desconozco', 'mayores', 'antecedentes', 'senal', 'durmiendo', 'escucho', 'ruido', 'patio', 'asom', 'vio', 'constanza', 'herrera', 'carvajal', 'saliendo', 'ventana', 'pieza', 'finalmente', 'quisiera', 'indicar', 'constanza', 'hija', 'pareja', 'don', 'miguel', 'cabrera', 'tambien', 'vecino', 'cite', 'desconozco', 'mayores', 'antecedentes', 'propiedad', 'conocemos', 'personalmente', 'senalar', 'varias', 'oportunidades', 'constanza', 'ido', 'habitacion', 'solicitar', 'diversas', 'cosas', 'limones', 'etc', 'jamas', 'dado', 'confianza', 'ingrese', 'asi', 'pieza', 'ultimo', 'indicar', 'segun', 'averigue', 'constanza', 'viviendo', 'cochrane', 'aconcagua', 'poblacion', 'trigal', 'calera', 'costado', 'numeracion', 'todd', 'cuanto', 'puedo', 'senalar', 'respecto'

Tabla 4.5: Ejemplo de una causa antes y después del preprocesamiento

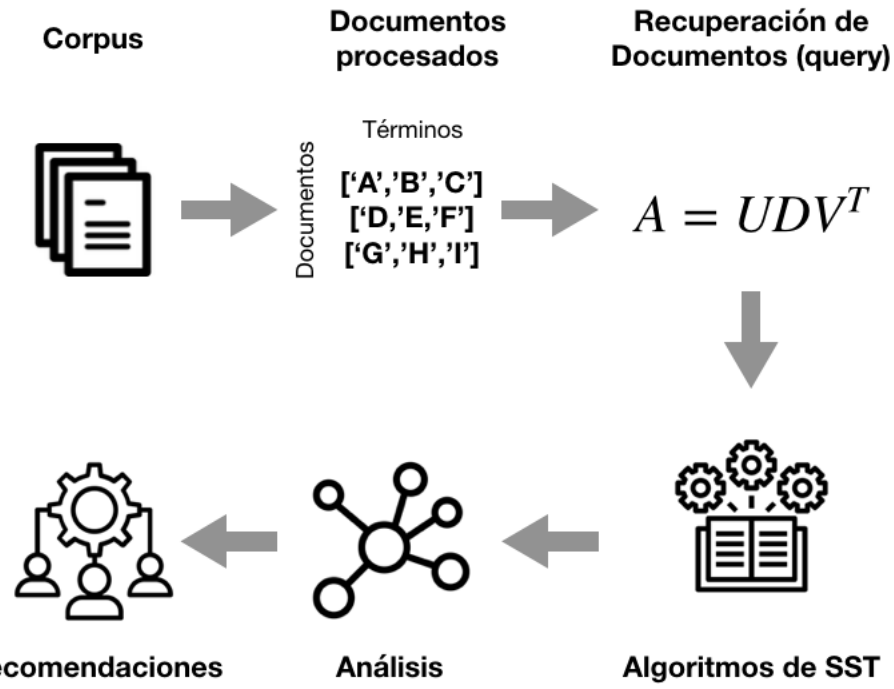


Figura 4.4: Descripción del proceso realizado para el desarrollo de recomendaciones al Ministerio Público

4.4.1. Recuperación de documentos

Para la recuperación de documentos se implementará el algoritmo Latent Semanctic Indexing (LSI) a través de la librería gensim[5] en Python.

El primer paso para desarrollar un modelo de recuperación de documentos por query es especificar la query. Esta query se especificará según el objetivo o los documentos que se deseen recuperar (en este caso, los documentos de un Foco de Investigación). Se derivarán los términos entonces a partir de una análisis exploratorio de los datos que se desea caracterizar.

La eficiencia del proceso de recuperación de documentos se evaluará en función de métricas utilizadas para dicho fin, en este caso: Precision, Recall y F1-Measure. Cómo fue señalado en capítulo 3, cada una de estas métricas entrega información respecto a un aspecto de la búsqueda por queries, siendo F1-Measure la métrica que balancea los resultados de Precision y Recall. No obstante, no siempre es posible guiarse exclusivamente por estas métricas, ya que pueden existir condiciones específicas de la búsqueda que se establezcan cómo restricciones, ante lo cual se escogerán los valores posibles dentro del espacio factible de soluciones.

[5]La documentación de esta libreria para LSI puede ser encontrada en: https://radimrehurek.com/gensim/models/lsimodel.html.

4.4.2. Computo de similitud semántica textual

En el trabajo de Alvarez (2017), se realiza un análisis comparativo de los algortimos Doc2Vec, Doc2VecC, WMD y Sent2Vec en una tarea de similitud semántica textual aplicado sobre un corpus de publicaciones científicas. Los resultados se muestran en la Tabla 4.6.

	Baseline	Doc2Vec	Doc2VecC	WMD	Sent2Vec
Títulos	0.91	0.65 (1M)	0.87 (1M)	0.90	0.91 (1M)
Abstracts	0.93	0.86 (1M)	0.92 (50K)	0.92	0.87 (100K)
Cuerpos	0.96	0.97 (500K)	0.94 (10K)	–	0.83 (10K)

Tabla 4.6: Resultados del trabajo de Alvarez (2017). Incluye la cantidad de muestras de entrenamiento entre paréntesis cuando es relevante.

Cómo se ha descrito anteriormente, en este trabajo se evaluará la pertinencia de utilizar la similitud semántica textual para la vinculación criminal, para lo cual se evaluaron los algoritmos WMD y Doc2Vec. Es importante destacar, que en nuestro sistema evaluaremos la similitud semántica textual dentro de los documentos que han sido recuperados por LSI, permitiendo aliviar la carga computacional del sistema y haciendo foco en aquellas causas que ya poseen una similitud en la frecuencia relativa a campos de búsqueda (una query).

4.4.3. Implementación Word Mover's Distance

Para la implementación de Word Mover's Distance (WMD), se utilizó la implementación de la libreria gensim[6] (Adim Řehůřek y Petr Sojka, 2010) en Python.

Tal como se señaló en el capítulo anterior, WMD resuelve un problema un problema de transporte mediante programación lineal, donde el espacio vectorial sobre el cual se resuelve este problema está dado por la representación que entregan los Word embeddings, por tanto, para computar una medida de similitud entre documentos, es necesario definir primero los embeddings de estos documentos.

4.4.4. Implementación Word embeddings

Jorge Pérez[7] computó sobre el Spanish Billion Word Corpus project[8], los word embeddings utilizando los algoritmos FastText y Glove. Estos embeddings fueron descargados desde el GitHub Uchile-NLP. En la misma página se encuentran disponibles los embeddings computados sobre el mismo corpus de 1.5 billones de palabras utilizando el algoritmo Word2Vec, el

[6]La documentación de esta libreria para WMD puede ser encontrada en: https://radimrehurek.com/gensim/models/keyedvectors.html.

[7]Doctor en Ciencias de la Ingeniería y académico del departamento de Ciencias de la Computación de la Universidad de Chile.

[8]Spanish Billion Word Corpus and Embeddings: http://crscardellino.me/SBWCE/.

cual fue computado por Cristian Cardellino[9].

Dado lo anterior, no fue necesario implementar nada para computar estos embedding. De todas formas, cada uno de estos algoritmos cuenta con una versión abierta y pública del código que se utilizó para su desarrollo[10].

4.4.5. Implementación Doc2Vec

Para la implementación de Doc2Vec, también se utilizó la librería gensim[11], utilizando el modelo de 'memoria distribuida' (PD-DM), donde se escogio un tamaño de los vectores de características de 300×1, ya que en la práctica ha demostrado ser el con mejor desempeño en tareas de clasificación y *analogy task*. También se ignoraron aquellas palabras con frecuencia 1 en el corpus, una tasa de aprendizaje inicial (α) de 0,5 y un contexto de 50 palabras para cada palabra objetivo.

4.4.6. Implementación Latent Dirichlet Allocation

Para la implementación de Latent Dichlet Allocation (LDA), una vez más se utilizó la librería gensim[12]. Una desventaja de esta librería es que algunos parámetros como α y β no pueden ser modificados. Sin embargo, se ha escogido esta libreria por sobre otras implementaciones por la gran integración que posee con la libreria pyLDAvis, cuya implementación y objetivo en este trabajo se discuten en la sección 4.5.4.

Por otro lado, se ha demostrado empíricamente que para una baja cantidad de tópicos, los resultados varían poco en términos de Perplexity calculada por el modelo según el parámetro α[13], lo que en la práctica significa que la poca flexibilidad de este parámetro en la implemenación escogida no significan un problema relevante, pues dada las restricciones operativas del usuario final de nuestra herramienta, no es posible lidiar con una alta cantidad de tópicos[14].

[9] Al momento de escribir este trabajo, Cristian era estudiante del doctorado de Ciencias de la Computación de la Universidad Nacional de Córdoba.

[10] En el caso de Word2Vec, el código se encuentra disponible en: https://code.google.com/archive/p/word2vec/. Para GloVe: https://nlp.stanford.edu/projects/glove/, y para FastText: https://github.com/facebookresearch/fastText.

[11] La documentación de esta libreria para Doc2Vec puede ser encontrada en: https://radimrehurek.com/gensim/models/doc2vec.html.

[12] La documentación de esta libreria para LDA se puede encontrar en https://radimrehurek.com/gensim/models/ldamodel.html.

[13] Es el mismo creador de Gensim discute en su blog a través de un ejercicio cómo varían los resultados según para diferentes configuraciones de α. Los resultados se pueden obserbar en: https://rare-technologies.com/python-lda-in-gensim-christmas-edition/.

[14] En este caso, consideraremos que 40 o más tópicos significan una cantidad alta de tópicos

4.5. Evaluación de resultados

4.5.1. Implementación de visualizaciones y resumen de contenido

El último objetivo específico de este trabajo consiste en desarrollar visualizaciones que permitan entender la información contenida en los conjuntos de recomendación, los cuales se obtienen a través de los 3 pasos señalados anteriormente: recuperación de documentos en base a queries, filtrar los documentos recuperados en base a su similitud semántica textual y la búsqueda de tópicos o conjuntos de causas que posean una alta similitud.

Para lograr el desarrollo de una herramienta que sea no sólo explicativa, sino que también entendible y flexible para un usuario final es que se han desarrollado 2 mecanismos para comprender los resultados:

- Implementación de visualizaciones que permitan representar la pertenencia de determinados documentos a estructuras naturales que se han encontrado en los documentos dado su similitud. Estás estructuras pueden estar determinadas por la similitud semántica textual entre los documentos o tópicos encontrados a través de Latent Dirichlet Allocation.

- Aglomeración de causas representados por los términos más frecuentes (keywords) para los conjuntos que se han encontrado. Además de visualizar las estructuras que se dibujan por el proceso de filtro y agrupación de causas descrito anteriormente, es útil mostrar cuales son las palabras más frecuentes o con mayor relevancia dentro de los conjuntos encontrados, de esta forma, se espera comunicar el contenido que vincula a un determinado conjunto de causas, permitiendo al usuario final ser parte del proceso de toma de decisión sobre que conjunto de causas investigar.

4.5.2. Representación de los documentos a través de grafos

El objetivo de implementar un grafo que muestre las relaciones entre las causas luego de haber computado y filtrado los arcos con menor grado de cercanía, es mostrar al usuario final la topología de una red que grafica las afinidades y diferencias entre causas. Así, al modificar el criterio de filtro para las causas (el cual se encuentra en [0,1]) se puede ver como la cantidad de causas en el grafo disminuye y las relaciones entre estas se van desagregando y se generan estructuras naturales de causas que están conectadas entre si y no con el resto (lo cual podría ser entendido como una comunidad o cluster de causas).

Para esto se utilizó la librería de Python Networkx[15] y el algoritmo Force Atlas 2[16] (Jacomy, M et. al, 2014), el cual dibuja la topología de la red basado sólo en información contenida por la red simulando fuerzas de repulsión entre los nodos. El resultado es un grafo no dirigido en donde causas con mayor similitud se encuentran más cerca y alejadas de causas

[15]Networkx: https://networkx.github.io/.
[16]ForceAtlas 2: https://github.com/bhargavchippada/forceatlas2.

Conjunto	Nº de causas	Keywords
Conjunto 1	9	poblacion, vehiculo, luego, individuos, automovil, cargo, procedieron, cuales, valparaiso, telefono
Conjunto 2	5	domicilio, guardia, parte, avaluo, camaras, individuos, seguros, espera, anos, testigos

Tabla 4.7: Ejemplo de una tabla de resumen para conjuntos de causas

con las cuales no se encuentran conectadas.

4.5.3. Tablas de resumen de contenido

Cada conjunto de causas que es generado por el proceso desarrollado es representado por las palabras más frecuentes dentro de ese conjunto de causas, eliminando las palabras más frecuentes dentro de todo el corpus y las palabras que han sido utilizadas como keywords en la búsqueda de documentos. La eliminación de estas palabras se realiza con el objetivo de restar información común a todas las causas, permitiendo llegar a palabras que son más relevantes sólo dentro de un determinado conjunto de causas.

En la Tabla 4.7 se aprecia un ejemplo de una tabla de resumen para 2 conjuntos de causas y los keywords que representan esos conjuntos. En este caso, se ha elegido representar a los conjuntos con las 10 palabras más frecuentes, eliminando las 10 palabras más comunes de todo el corpus y los términos para la búsqueda de documentos.

4.5.4. Implementación de una visualización interactiva para tópicos

La representación de clusters de documentos a través de términos frecuentes suelen ser una poderosa herramienta para sintetizar aspectos claves de las causas que se encuentran en un conjunto. No obstante, cuando estos conjuntos son muy grandes, es cada vez mas probable que exista heterogeneidad en los contenidos dentro de esos conjuntos, al mismo tiempo que 10 (o más) palabras claves pierden poder explicativo. En dichos casos se ha optado por la implementación de Latent Dirichlet Allocation (LDA) como una herramienta para lidiar con dicha heterogeneidad.

No obstante, aún con la implementación de LDA es necesario poder responder las siguientes preguntas: 1. ¿De qué trata cada tópico? 2. ¿Cuán prevalente es cada tópico? y 3. ¿Cómo se relacionan estos tópicos?. Para responder de forma simple esas preguntas se implementó con la ayuda de la libreria de Python pyLDAvis[17] la herramiento LDAvis (Sievert y Shirley, 2014), la cual consiste en una visualización dinámica basada en D3[18] con 2 elementos principales:

[17]pyLDAvis: https://github.com/bmabey/pyLDAvis.

[18]D3.js (o simplemente D3 por las siglas de Data-Driven Documents) es una librería de JavaScript para producir visualizaciones dinámicas de datos en navegadores web.

1. Una visualización que presenta una vista global del modelo de tópicos y responde las preguntas 2 y 3. En esta vista, se representan los tópicos como círculos en un plano bidimensional cuyos centros se determinan calculando la distancia entre los temas, y luego mediante el uso de escalamiento multidimensional se proyecta las distancias intertópicos en dos dimensiones.
2. Una visualización de barras que representan los términos individuales (palabras) que son más útiles para interpretar un determinado tópico, y permite a los usuarios responder a la pregunta 1 (¿De qué trata cada tópico?). Un par de barras superpuestas representan tanto la frecuencia de todo un corpus de un término determinado, así cómo la frecuencia específica del termino en un determinado tópico.

Con la ayuda de esta visualización se pretende facilitar el análisis de conjuntos con 50 o más causas[19] para encontrar subconjuntos que posean mayor homogeneidad en su contenido/relato y permita realizar la investigación conjunta de estas causas (que es el objetivo final).

4.5.5. Juicio de expertos

El juicio de expertos se define como una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones. La identificación de las personas que formarán parte del juicio de expertos es una parte crítica en este proceso, frente a lo cual Skjong y Wentworht (2001) proponen los siguientes criterios de selección: (a) Experiencia en la realización de juicios y toma de decisiones basada en evidencia o experticia (grados, investigaciones, publicaciones, posición, experiencia y premios entre otras), (b) reputación en la comunidad, (c) disponibilidad y motivación para participar, y (d) imparcialidad y cualidades inherentes como confianza en sí mismo y adaptabilidad. Utkin (2006) plantea que el juicio de expertos en muchas áreas es una parte importante de la información cuando las observaciones experimentales están limitadas, lo cuál es bastante similar al escenario que enfrentamos.

Dado que este problema no trata de uno de los clásicos problemas de clasificación, regresión o asociación, es que debemos definir otra métrica de desempeño. En nuestro caso estamos interesados en conocer cuál es el nivel de similitud en la comisión de un delito, caracterizado por su *modus operandi* o la descripción de los sujetos que cometen un delito. No obstante, lo segundo, es decir, la descripción de los delincuentes suele ser difusa, incompleta y en muchos casos puede no existir. Por otro lado, es mucho más probable que los algoritmos de similitud semántica textual capturen la similitud medida en términos de la similitud del *modus operandi*, ya que en la denuncia lo que se suele describir son los hechos que dieron origen al delito.

[19]No existe un consenso respecto al número mínimo de palabras que debe posee un corpus para que los resultados sean válidos, sin embargo, de nuestro conocimiento el corpus de menos tamaño sobre el que se han reportado resultados 'válidos' para LDA es de 10.470 palabras (Crossley et. al, 2018). Además en el mismo trabajo se demuestra empíricamente que los resultados para LDA mejoran en la medida en que aumenta el tamaño del corpus sobre el cual se emplean el algoritmo, lo cual cobra sentido en LDA ya que es un algoritmo que converge mediante la estimación de máxima verosimilitud y un largo subóptimo del corpus podría dificultar la labor de convergencia a un óptimo global o al menos uno que brinde resultados aceptables.

Objetivo de la validación	Verificar si causas que no están asociadas al mismo foco de investigación poseen un *modus operandi* semejante.
Expertos	Analistas del Ministerio Público encargados de contituir el Foco de Investigación en seguimiento.
Método de validación	Un experto analiza cada vez una causa que pertenece a un foco en donde el imputado es desconocido y 10 causas que no se encuentran en el foco para determinar si dichas causas pueden estar vinculadas o poseen el mismo *modus operandi*.

Tabla 4.8: Principales aspectos que se tuvieron en cuenta para determinar el proceso de validación por juicio experto

Finalmente, para nuestra evaluación de los resultados se buscará validar que causas con un elevado nivel de similitud semántica textual poseen un *modus operandi* similar. Para esto, se consultará a los analistas del Ministerio Público responsables de constituir el Foco de Investigación Nº23 de Valparaíso si un conjunto de causas seleccionadas por el mecanismo de búsqueda y asociación descrito poseen el mismo método de comisión que las causas que ya están en el foco y por tanto pueden, ser incorporadas dentro del foco. Este ejercicio es denominado en la Fiscalía 'Seguimiento del Foco', es decir, evaluar si han surgido nuevas causas que puedan ser incorporadas en un foco ya existente.

En la Tabla 4.8 se resumen los principales aspectos que se tuvieron en cuenta para determinar el proceso de validación a través de juicio experto sobre la similitud entre las causas del Foco de Investigación Nº23 de Valparaíso y un listado de las 10 causas más similares determinadas por lo algoritmos de similitud semántica textual.

Capítulo 5

Resultados

El despliegue de resultados se dividirá en 2 partes: la primera parte reportará el proceso completo desde la limpieza de los datos, la búsqueda de documentos a través de una query, la agrupación de causas en base a su similitud semántica textual y la visualización y/o reporte de los resultado. Los resultados de esta parte corresponden al ejercicio de generar un conjunto de causas lo más acotado posible a partir de las 3.803 causas descritas anteriormente, que contengan en la medida de lo posible, al Foco Investigativo Nº23 de la Fiscalía de San Antonio. En paralelo se mostrará cómo evoluciona la recuperación de las causas del Foco de Investigación. Es importante destacar que a través de este proceso se construye una matriz de similitudes entre todas las causas, por lo que al finalizar este proceso, es posible elaborar un ranking de las causas más similares a las contenidas en el Foco de Investigación, pero que no se encuentran incluidas en él, lo cual representa un espacio de oportunidades.

En la segunda parte de este capítulo se presentarán las conclusiones y validaciones por juicio experto del ejercicio de seguimiento del Foco de Investigación. Esto quiere decir, que a través de la validación de analistas del Ministerio Público se intentará incluir causas al foco ya descrito a través de una priorización a partir de la similitud entre causas computada por los algoritmos. En este caso, se entregó un listado de las 10 causas más similares a aquellos delitos en el Foco de Investigación donde el autor es desconocido y se encuentra sin captura (o al menos era desconocido y se encontraba sin captura a la fecha en que se realizó este ejercicio).

Para resumir: en la primera parte se reportan los resultados de la asociación de causas desde la búsqueda por query, lo cual a nuestro juicio, constituye el principal aporte para las labores de investigación del Ministerio Público. En la segunda parte se da seguimiento al Foco de Investigación Nº23 de San Antonio, validando por expertos la pertinencia de incluir un conjunto de causas al Foco de Investigación.

(a) Corpus completo

(b) Foco de Investigación

Figura 5.1: Nubles de palabras para los relatos del corpus completo y el Foco de Investigación

5.1. Asociación de causas por similitud semántica textual

5.1.1. Análisis exploratorio de los datos

Como se señaló en el capítulo anterior, el Ministerio Público ha brindado 2 conjuntos de datos: el primero cuenta con 3.803 registros de delitos cometidos en la Macrozona de San Antonio, donde se encuentran robos con violencia e intimidación, así como robos en lugar habitado o destinado a la habitación. El segundo conjunto de datos es el Foco de Investigación Nº23 que consiste de 16 causas que se investigan de manera conjunta, todos de robos en lugar habitado o destinado a la habitación y cuya caracterización en término de tipos de crímenes y especies sustraídas ha sido presentado con anterioridad.

En la Figura 5.1 se pueden apreciar dos Word Clouds, uno para el conjunto de 3.803 causas y otro para el Foco de Investigación. Los Words Clouds son una de las técnicas de visualización de datos textuales más comunes, cuyo objetivo es brindar información global sobre el contenido del conjunto de datos. En los Word Clouds, cada palabra se representa a sí misma y el tamaño de la fuente representa la frecuencia de dicha palabra dentro del conjunto, así las palabras más frecuentes en el conjunto de datos se visualizarán con un tamaño de letra mayor. Algo interesante que se puede apreciar en la Figura 5.1 es que en el conjunto de causas del Foco de Investigación aparecen como términos frecuentes las palabras `sujeto`,`inmueble` y `detenido`, términos que caracterizan el tipo de delitos con mayor proporción en este conjunto.

En la Tabla 5.1 se pueden apreciar los 10 términos más frecuentes normalizado cada 10.000 palabras[1] en donde se hacen más evidentes las diferencias entre ambos conjuntos de causas. Si bien comparten términos frecuentes estos corpus, hay términos donde difieren, los cuales nos permitirán construir una query que incluya dichos términos ya que son una buena fuente de diferenciación entre ambos conjuntos.

En la Tabla 5.2 se pueden apreciar los 10 N-grams más frecuentes para ambos corpus. Los N-grams se definen como una subsecuencia de N elementos de una secuencia determinada. En este caso, los N-grams serán una subsecuencia de palabras de la secuencia de palabras

[1]Esta normalización quiere decir que la cifra que aparece en el tabla representa la cantidad de veces que aparece dicho término cada 10.000 palabras. El objetivo de esta normalización es permitir la comparación de la frecuencia de palabras para corpus de diferentes dimensiones.

Conjunto completo de causas		Foco de Investigación	
Término	Frecuencia cada 10.000 palabras	Término	Frecuencia cada 10.000 palabras
marca	103	personal	106
lugar	100	lugar	106
especies	85	especies	98
personal	80	identidad	87
local	79	inmueble	73
victima	76	domicilio	73
domicilio	76	detenido	68
parte	73	local	60
denunciante	68	victima	56
guardia	67	lesiones	53

Tabla 5.1: Términos más frecuentes normalizados cada 10.000 palabras para ambos corpus de documentos

Conjunto completo de causas		Foco de Investigación	
N-grams	Conteo absoluto	N-grams	Conteo absoluto
quedo, espera, parte, local	1265	denunciante, quedo, espera, parte	5
denunciante, quedo, espera, parte	1038	quedo, espera, parte, local	5
victima, quedo, espera, parte	538	senoret, esquina, avda, errazuriz	3
anos, chileno, soltero, estudios	490	almirante, senoret, esquina, avda	3
especies, mas, abajo, detallan	422	comunicado, radial, central, comunicaciones	3
direccion, desconocida, especies, sustraidas	395	individuos, desconocidos, habian, ingresado	3
lugar, direccion, desconocida, especies	370	trasladado, sar, valparaiso, medico	2
chileno, soltero, estudios, medios	368	ruta, lote, pachacamita, calera	2
celular, marca, samsung, modelo	330	corta, pluma, marca, stainless	2
huir, lugar, direccion, desconocida	321	manifiesta, tener, sospechas, determinadas	2

Tabla 5.2: N-grams más frecuentes para ambos corpus de documentos

que constituyen una denuncia. El objetivo de buscar los N-grams más frecuentes es relevar si existen frases o descripciones con alta frecuencia que puedan brindar información relevante. Para ambos corpus, los 2 N-grams más frecuentes tienen relación con una estructura común de las denuncias que tiene relación con una formalidad que se debe establecer en los relatos, en este caso, que la víctima queda a la espera de la citación de la Fiscalía local. Existen otros N-grams que permiten relevar especies que han sido sustraídas de forma frecuente cómo `celular, marca, samsung, modelo`. Finalmente, existe un N-gram que llama la atención: `almirante, senoret, esquina, avda, errazuriz`. Este N-gram hace referencia a una ubicación geográfica, y dado que el contexto de estos documentos son delitos, podría revelar la ubicación de un *hot spot*.

Finalmente, luego de analizar las diferencias en los términos que constituyen los relatos de ambos corpus, algo interesante de analizar es si existen diferencias en la extensión de

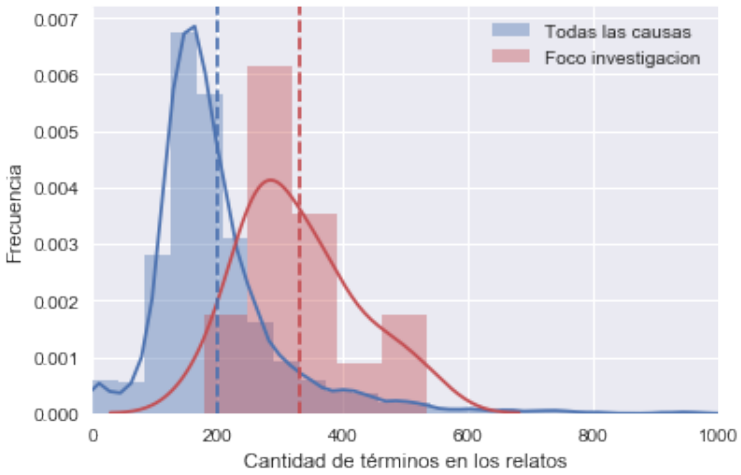


Figura 5.2: Distribución de la cantidad de términos por corpus

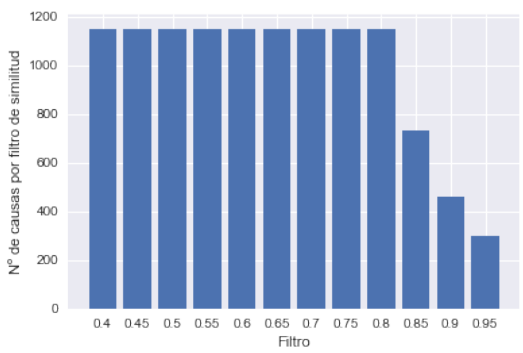
los relatos de ambos conjuntos. En la Figura 5.2 se puede apreciar que existen diferencias significativas en la cantidad de palabras utilizadas en las denuncias de ambos conjuntos. Luego de haber preprocesado los documentos de ambos corpus (lo que entre otras cosas implica eliminar *stop words*), las denuncias de todos los documentos tienen una extensión en promedio de 198 palabras, mientras que en las denuncias del Foco de Investigación, estos relatos poseen una extensión en promedio de 330 palabras, es decir son en promedio un 51,5 % más extensos los relatos del Foco de Investigación. Algunas hipótesis al respecto tienen que ver con la cantidad de información contenida en ambos corpus, ya que en el Foco de Investigación existen causas que han sido agrupadas en ese foco dada cantidad de información disponible en esa causa y su naturaleza. Así mismo, es mucho más probable que muchas de las causas que no están en el foco sean denuncias con muy pocos antecedentes sobre la comisión del delito o del delincuente.

5.1.2. Elección de los términos de búsqueda

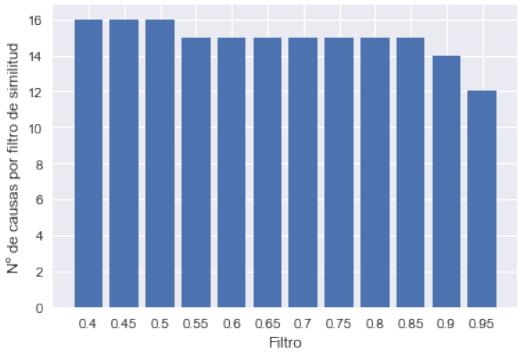
El primero objetivo específico de este trabajo consiste en la implementación de algoritmos de recuperación de documentos, en este caso Latent Semantic Indexing. Para esto es necesario definir cual será la consulta que se realizará sobre el corpus de documentos.

En la Figura 5.1 se puede apreciar que para los relatos de ambos conjuntos (el corpus completo y el Foco Investigativo) existen diferencias como la aparición de las palabras `inmueble, detenido e individuo/sujeto` en los relatos del Foco de Investigación, lo cual se puede explicar por las razones señaladas anteriormente, es decir, que el Foco de Investigación registra causas sobre robos en lugar habitado o destinado a la habitación donde se encuentra sobrerrepresentada la cantidad de causas donde el delincuente fue detenido tras o durante la comisión del delito. Dado lo señalado anteriormente se ha escogido incorporar los siguientes términos en la query:

```
Query = ['detenido', 'captura', 'lesiones', 'inmueble', 'identidad', 'domicilio']
```



(a) Recuperación de documentos en el corpus completo



(b) Recuperación de documentos en el Foco de Investigación

Figura 5.3: Cantidad de documentos recuperados según similitud con la query

5.1.3. Recuperación de documentos

Cómo se describio en el capítulo 3, Latent Semantic Indexing recupera los documentos que más similitud poseen con un query, lo cual se computa mediante la similitud coseno entre la representación vectorial de la query y la representación vectorial de los documentos en el nuevo espacio k-dimensional $(cos(\hat{\mathrm{d}}_\mathrm{i}, q))$ que genera el modelo. Esto implica que se debe escoger un punto de corte, en el cual se considera que a partir de ese punto los documentos serán recuperados. Dado que se computa mediante similitud coseno, esta métrica se encuentra en el conjunto $[0, 1]$.

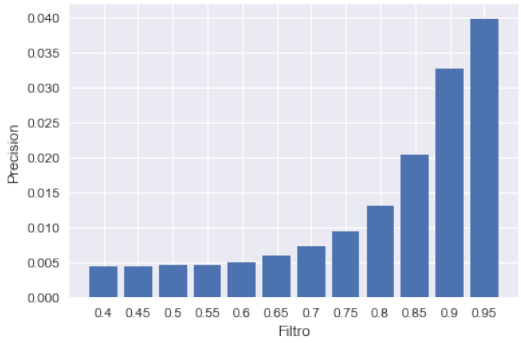
En la Figura 5.3 se puede apreciar cómo decae la cantidad de causas que cumplen con la condición de ser mayores a un determinado valor de similitud. De esta Figura podemos concluir que la query ha caracterizado bastante bien a las causas del Foco de Investigación ya que las causas del Foco de Investigación son recuperadas en una proporción mucho mayor a las del corpus completo a medidas de similitud más cercanas a 1.

Recordemos que nuestro objetivo es maximizar el ratio:

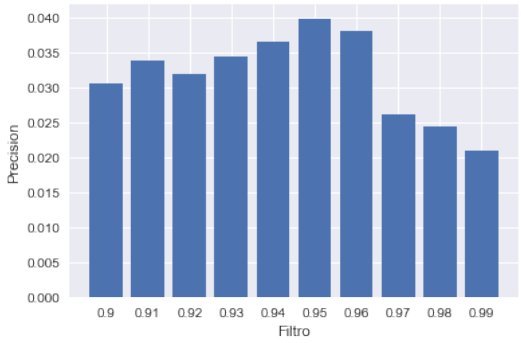
$$\frac{n(\text{Causas recuperadas del Foco de Investigación})}{n(\text{Causas recuperadas del corpus completo})} \tag{5.1}$$

Donde $n(\cdot)$ representa el cardinal de un conjunto.

Analizaremos las métrica de desempeño en términos de recuperación de documentos para estudiar dónde se encuentra el punto óptimo para setear un nivel de similitud. En la Figura 5.4 se muestra a la izquierda cómo varía el valor de Precision de acuerdo al punto de similitud utilizando los mismos valores que en la Figura 5.3. No obstante, es posible observar que el valor óptimo probablemente se encuentre en en el intercalo [0,9, 1,0], que es precisamente lo que muestra el lado derecho de la Figura 5.4. Es posible concluir entonces que el valor máximo para Precisión se encuentra para el valor 0,95 de similitud entre query y documentos, de la misma forma la Figura 5.5 nos indica que el valor máximo para F1-Measure (que es la

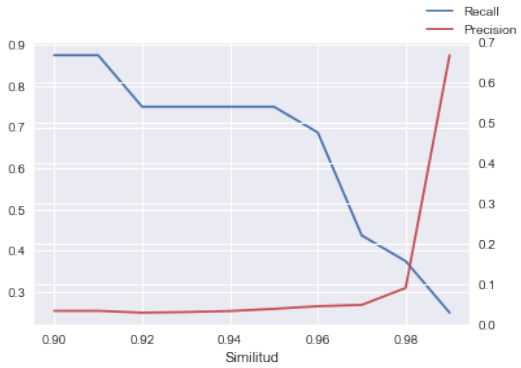


(a) Precision en función de valor de similitud para el rango (0,4, 0,95)

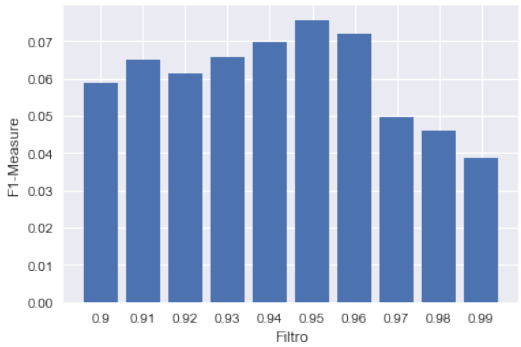


(b) Precision en función de valor de similitud (0,90, 0,99)

Figura 5.4: Precision en función de valor de similitud



(a) Precision y Recall por nivel de similitud



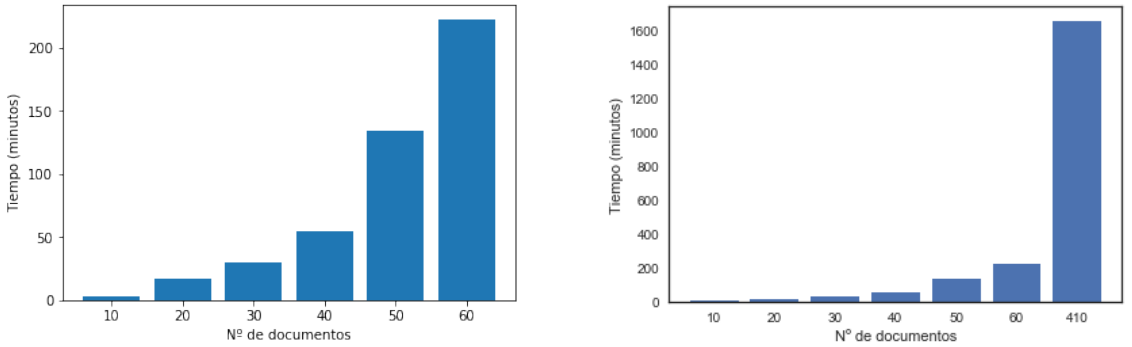
(b) F1-Measure por nivel de similitud

Figura 5.5: Precision, Recall y F1-Measure por nivel de similitud

mñetrica que balancea Recall y Precision), alcanza su máximo en 0,95. No obstante, dado que la muestra de documentos que queremos recuperar es muy acotada (tan sólo 16 documentos), es que hemos definido tener un valor mínimo de Recall tal que: Recall $\geqslant$ 80 %. El 80 % de nuestras causas corresponde a 13 documentos y el valor máximo para el que se alcanza dicho Recall es para una similitud de 0,91. En la Figura 5.5 se puede apreciar en el eje izquierdo los valores de Recall y en el eje derecho los valores de Precision. Así, el Recall para una similitud de 0,91 es de un 88 %.

Finalmente, con el objetivo de minimizar causas ruidosas, y buscar aquellos documentos que mayor similitud tengan a las causas del Foco de Investigación es que se han filtrado todas las causas que luego de su procesamiento tengan menos de 50 palabras. Esto, ya que como se ha visto en la Figura 5.2, las causas del Foco de Investigación poseen una extensión mayor en su cantidad de palabras. A continuación se resumen los parámetros y condiciones que se han escogido y las métricas resultantes de estas decisiones:

```
[length(doc)>50] & sim(query,doc)>0.91] = 414 causas(10.88% del corpus)
                    Recall = 88% & Precision = 3.38%
```



(a) Tiempo de computo (b) Proyección lineal del tiempo de computo

Figura 5.6: Tiempo de computo para el algoritmo WMD

5.1.4. Computo de similitud semántica textual: Word Mover's Distance

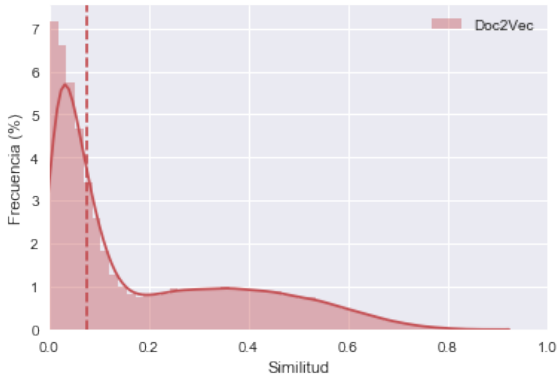
Una vez acotado el conjunto de causas a través de la recuperación de documentos por query, se puede realizar una búsqueda y asociación a través de similitud semántica textual. Nuestro primer intento fue a través del algoritmo Word Mover's Distance, el cual como fue descrito en el capítulo 3, corresponde a un proceso de optimización similar a minimizar la distancia total que implicaría transformar todas las palabras de un documento en las palabras de otro, utilizando como espacio vectorial la representación de Word Embeddings de cada término y por tanto documento.

Cómo se aprecia en la Figura 5.6 el tiempo de cómputo para este algoritmo crece de forma polinomial en la cantidad de causas a las que se busca computar su similitud. En el lado derecho de la Figura 5.6 se muestra un enfoque *naive* para la estimación del tiempo que llevaría computar la similitud entre todos los documentos que han sido recuperados por la búsqueda de query. Esto conllevaría un tiempo cercano a los 1600 minutos o 66.6 horas y esto es la cota inferior, ya que estima de forma cuasi lineal un tiempo que crece polinomial.

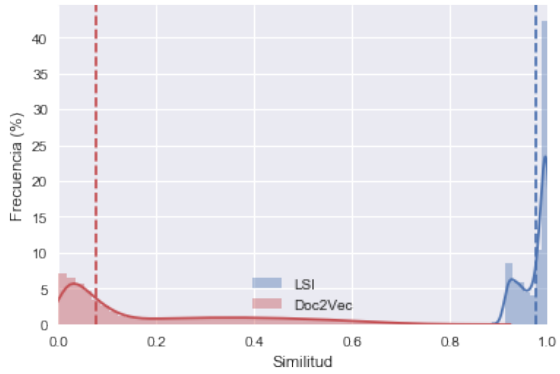
Lo anterior nos ha llevado a desistir nuestra intención de continuar la búsqueda de resultados a través de este documento. Esto ya que en la práctica, este conjunto de causas es bastante bajo e intentar desarrollar una solución a partir de esto no es replicable, ni escalable, ni justificable dada la existencia de otros modelos.

5.1.5. Computo de similitud semántica textual: Doc2Vec

Doc2Vec tiene la capacidad de ser entrenado fácilmente sobre el corpus de documentos que uno desea analizar, lo cual tiene la ventaja de que las palabras adquieren una representación según el dominio específico en el que se desea analizar las similitudes. Este algoritmo, generalmente es entrenado sobre corpus extensos dada su eficiencia, y por lo mismo es que se entrenó sobre el corpus completo, es decir, las 3.803 causas, en tan sólo 2.5 minutos.



(a) Distribución de la similitud computada por Doc2Vec



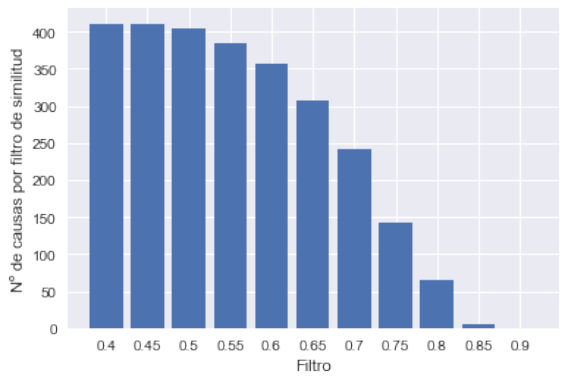
(b) Comparación de las similitudes según Doc2Vec y LSI

Figura 5.7: Comparación de las similitudes computadas por Doc2Vec y LSI

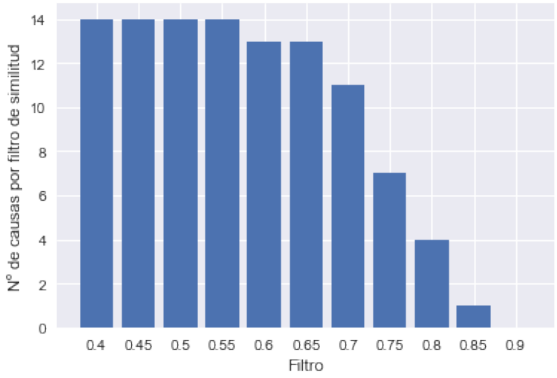
Dado que este corpus no es tan extenso, se han utilizado 500 iteraciones (es decir, el corpus es escaneado 500 veces), pese a que la literatura señala que con 200 iteraciones el modelo es capaz de converger a un óptimo. Se ha entrenado el modelo con una ventana (contexto) de 50 palabras para cada palabra target. Además se ha entrenado el modelo para que los vectores de salida sean de dimensiones 300×1 tal cómo se utiliza regularmente en la literatura. Luego de entrenar el modelo, se computa la matriz de distancias entre todos los documentos (la cuál posee dimensiones n^2), lo que requirió alrededor de 6 segundos.

En la Figura 5.7 se puede apreciar la radical diferencia que existe entre el cómputo de la similitud para el métdo LSI y Doc2Vec, siendo en promedio 0.07 el valor de la matriz de distancias que computa Doc2Vec y 0.97 la distancia promedio de las causas que fueron recuperadas de LSI. Estos métodos son radicalmente distintos en la forma en la que computan similitud y por tanto es difícil buscar una explicación certera para esta diferencia. Sin embargo, si podemos señalar con seguridad de que al ser LSI un modelo basado en el recuento de términos en común entre documentos, es más sensible a términos muy frecuentes en el corpus, por lo que podría estar basando su excesiva similitud entre las causas en estas estructuras comunes en las denuncias que ya hemos detectado a través de los N-grams y los términos más frecuentes.

En la Figura 5.8 se puede apreciar cómo evoluciona la recuperación de documentos según la similitud cálculada por Doc2Vec, esto para el corpus general y del Foco de Investigación recuperado por LSI. Además en la Figura 5.9 se puede apreciar cómo evoluciona Precision y el Recall según diferentes métricas de similitud. En conjunto las Figuras 5.8 y 5.9 nos entregan una visión global de cómo evolucionan la cantidad de documentos recuperados según el nivel de similitud estipulado. En la Figura 5.9 se puede apreciar claramente que existe un *trade-off* entre Recall y Precision, sin embargo, tal como se señaló en el Capítulo 3, en este caso, lo que corresponde maximizar es el F1-Measure, el cual alcanza su máximo para un valor de similitud en 0,8.

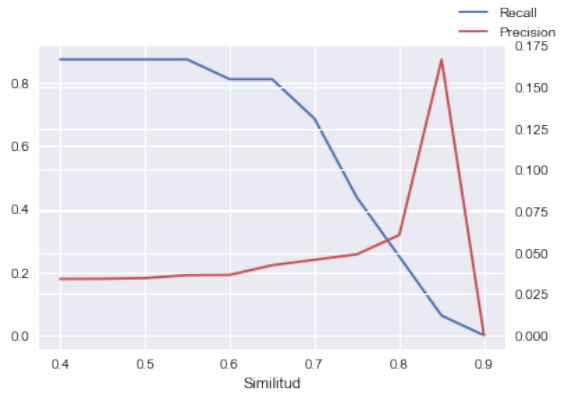


(a) Documentos recuperados del nuevo corpus por nivel de similitud

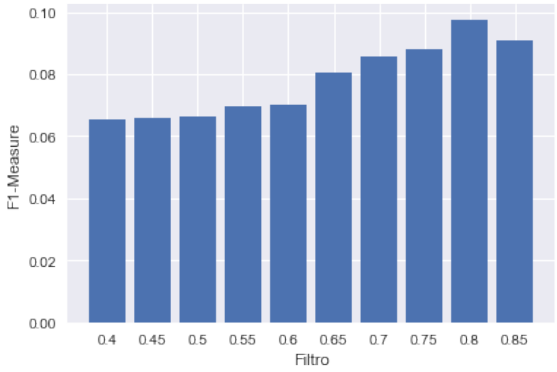


(b) Documentos recuperados del Foco de Investigación por nivel de similitud

Figura 5.8: Recuperación de documentos en las causas recuperadas por LSI del corpus completos y Foco de Investigación



(a) Documentos recuperados del nuevo corpus por nivel de similitud



(b) F1-Mesure por nivel de similitud

Figura 5.9: Métricas de recuperación de información por nivel de similitud

5.1.6. Interpretación de los resultados

Es necesario recordar que este trabajo no puede ser esclavo de las métricas de desempeño, ya que nuestro objetivo es agrupar causas que sean homogéneas en su contenido y hasta ahora no hemos vislumbrado cómo dar solución a aquello. En la sección anterior se mostraron múltiples métricas para determinar cual sería eventualmente el mejor valor en términos de similitud para agrupar causas. De aquí en adelante intentaremos explorar cómo se reflejan esas métricas de similitud en el contenido de las conjuntos de causas generadas.

Topografía de un grafo de documentos

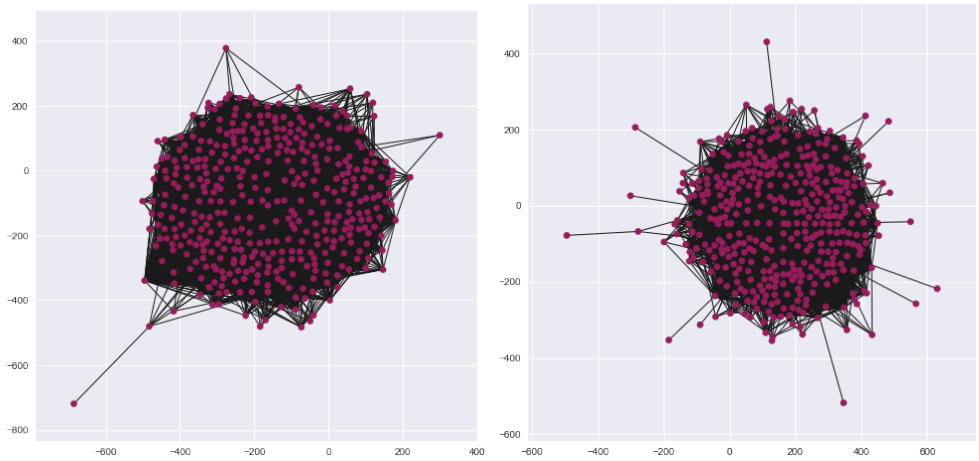
Dado que Doc2Vec genera una matriz de distancia entre los documentos, podemos llevar esa información a un grafo para visualizar si existen clusters o aglomeraciones que puedan dar luces de un conjunto de causas que han sido agrupadas naturalmente dado su similitud semántica textual. Para esto utilizaremos el algoritmo Force Atlas 2, que ubica los puntos acorde a las coordenadas que minimizan la energía total de la simulación de un sistema, donde se simula que los nodos poseen carga eléctrica (repulsiva entre ellos) y los vértices simulan ser resortes cuya extensión inicial es proporcional a la distancia en la matriz original (fuerza atractiva). Así se alcanza una configuración que ha demostrado ser muy efectiva para encontrar comunidades en grafos (Cherven, 2013).

En la Figura 5.10 se puede apreciar la transición entre diferentes topologías de un grafo que representa a los documentos agrupados según la similitud computada por Doc2Vec. Al aumentar la similitud implica que aquellos nodos que estén conectados por una similitud inferior a la exigida, dejan de estar conectados, es decir, desaparece la relación entre dichos nodos. Uno de los beneficios de este análisis es que permite observar desde qué punto se empiezan a generar aglomeraciones de forma natural, es decir, conjunto de causas que generan una comunidad conectada sólo entre ellos y desconectadas de otras comunidades. Coincidencia o no, la similitud que mayor comunidades genera es en torno a una similitud de 0.8, que es justamente la similitud que maximiza el F1-Measure.

Otra de las ventajas de este enfoque es que permite visualizar para nuestro caso en donde se ubican las causas del Foco de Investigación, respecto al resto de las causas. ¿Están las causas del Foco de Investigación es una comunidad densamente poblada?¿Están las causas del foco conectadas entre si?¿Existen causas en el Foco de Investigación en pequeñas comunidades que releven posibles causas que asociar? En la Figura 5.11 se puede visualizar la disposición del grafo para una similitud de 0.8 y en color azul las causas del Foco de Investigación.

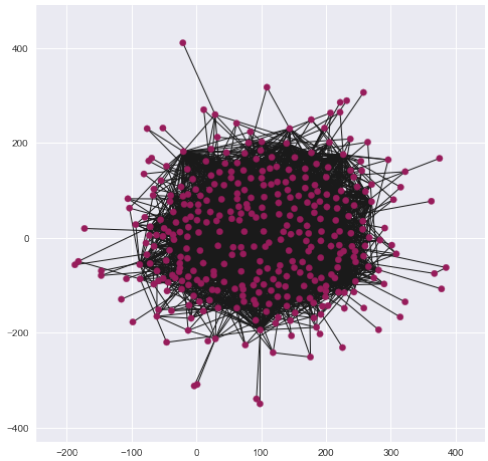
Word Clouds

De forma complementaria a analizar cómo evoluciona la transición entre diferentes topologías de un grafo que representa a los documentos, es de sumo interés comprender acerca de qué tratan los documentos en las comunidades o clusters que se generan de forma natural al modificar los niveles de similitud para la recuperación de documentos. Por ejemplo, en el

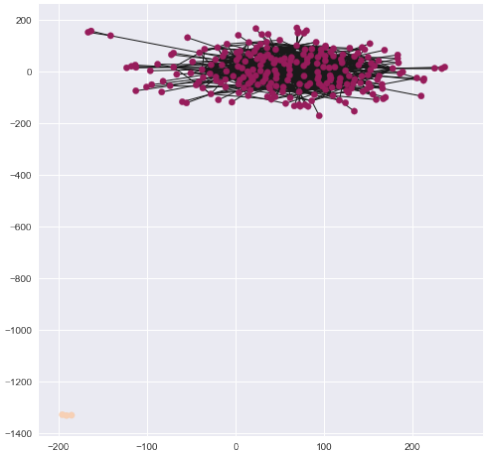


(a) Grafo para similitud = 0.4

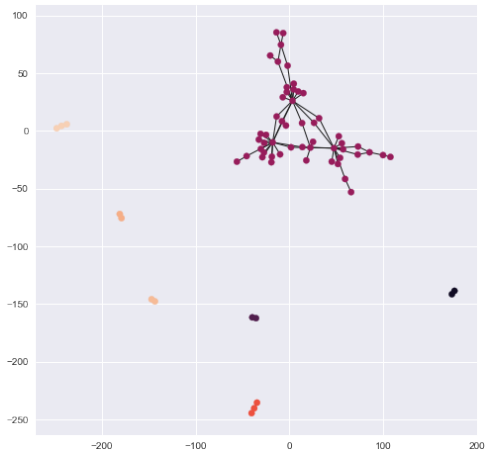
(b) Grafo para similitud = 0.5



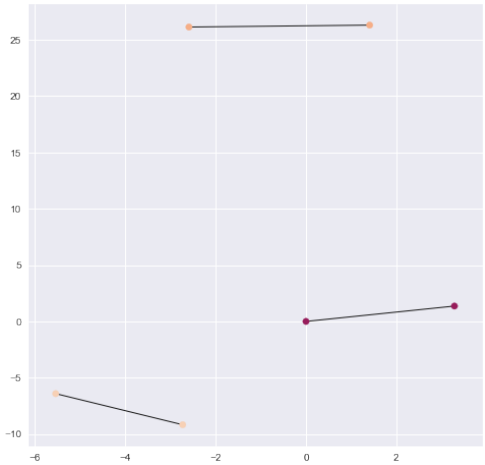
(c) Grafo para similitud = 0.6



(d) Grafo para similitud = 0.7



(e) Grafo para similitud = 0.8



(f) Grafo para similitud = 0.85

Figura 5.10: Cambios en la topología del grafo de documentos por nivel de similitud

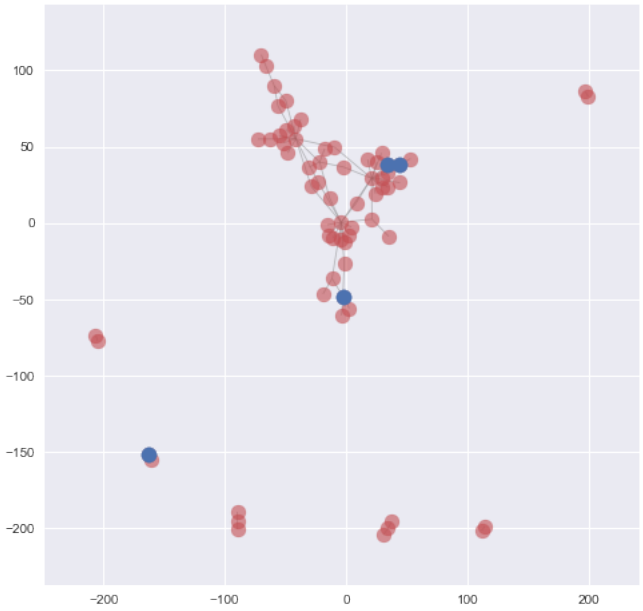


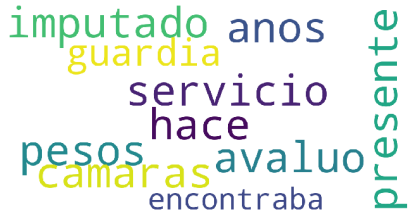
Figura 5.11: Visualización de la ubicación de las causas recuperadas del Foco de Investigación en el grafo

ejercicio anterior, en la Figura 5.11 se aprecia que para un nivel de similitud de 0,8, aparecen de forma natural 7 comunidades, es decir, subconjunto de causas del grafo que están más conectados entre sí que con el resto de la red. Una aproximación para resolver este problema es el de analizar los términos más relevantes dentro de los documentos de esas comunidades. La Figura 5.12 es un esfuerzo por comunicar dicha información de forma sencilla.

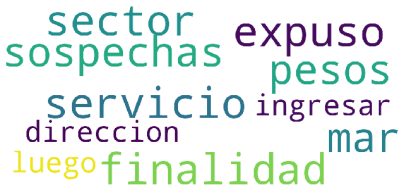
En la Figura 5.12 es posible apreciar que por ejemplo el cluster 7 hace referencia a una especia y un lugar (celular, cartagena), lo que eventualmente da cuenta de un problema de recurrencia. El cluster 3 da cuenta de un sector o calle (carampangue) y la insinuación de uno(s) detenido(s), lo cuál también podría brindar luces de un problema de recurrencia o fenómenos relacionados. Es importante destacar a su vez, que en los resultados de los Word Clouds han sido eliminados los términos más frecuentes en los documentos rescatados luego de la búsqueda por similitud computada por Word2Vec. A si mismo han sido eliminados los términos que se han utilizado en la query para el modelo de LSI con el fin de maximizar los términos que constituyen una diferencia en estos clusters.

Tablas de reporte general

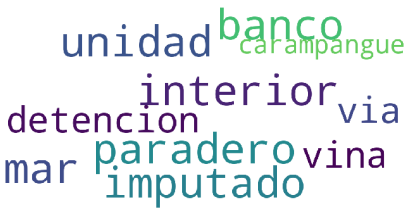
La visualización de los grafos en función de la similitud permite identificar comunidades y la dimensión de estas, lo que permite priorizar la búsqueda de causas relacionadas de acuerdo a los criterios de búsqueda, por ejemplo, "buscar una extensa comunidad de causas relacionadas con delitos a jóvenes los fines de semana por la noche"ó, "buscar una comunidad acotada y estrechamente vinculada a delitos con un *modus operandi* específico como el de un secuestro". Estos son ejemplos de delitos que podrían ser caracterizados a través de una query y cuya búsqueda de causas asociadas se podría representar como en la Figura 5.11



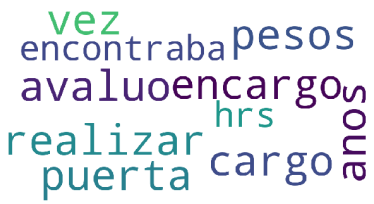
(a) Nube de palabras para cluster 1



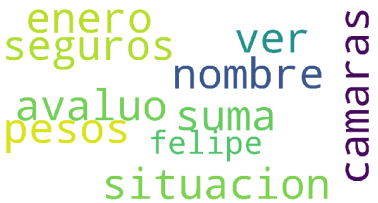
(b) Nube de palabras para cluster 2



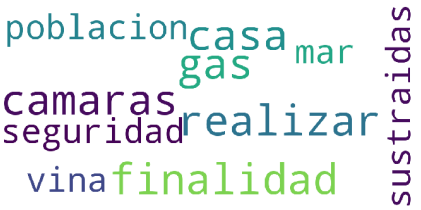
(c) Nube de palabras para cluster 3



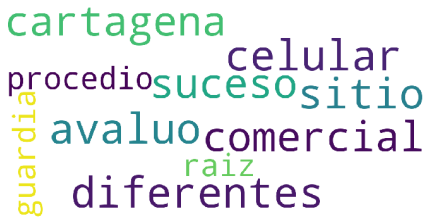
(d) Nube de palabras para cluster 4



(e) Nube de palabras para cluster 5



(f) Nube de palabras para cluster 6



(g) Nube de palabras para cluster 7

Figura 5.12: Word Clouds para los cluster identidicados

Cluster	Nº Causas	Causas del Foco	Keywords
cluster 1	52	3	guardia, pesos, encontraba, hace, presente, avaluo, servicio, camaras, espera, imputado
cluster 2	3	0	luego, sector, servicio, pesos, finalidad, sospechas, mar, expuso, direccion, ingresar
cluster 3	2	0	mar, vina, imputado, unidad, detencion, interior, banco, paradero, carampangue, achupallas
cluster 4	2	0	encontraba, puerta, avaluo, cargo, encargo, realizar, afectado, especialidad, gris, hecho
cluster 5	2	0	pesos, seguros, camaras, situacion, felipe, nombre, suma, avaluo, enero, comedor
cluster 6	3	0	mar, finalidad, casa, vina, poblacion, realizar, gas, camaras, seguridad, sustraidas
cluster 7	2	1	procedio, sitio, comercial, raiz, suceso, cartagena, levantamiento, celular, avaluo, guardia

Figura 5.13: Tabla de reporte general

donde es posible monitorear donde se ubican estos casos en una red de causas (en nuestro caso las causas de un Foco de Investigación).

Los Word Clouds complementan la información obtenida por la representación en red de los documentos, brindando información sobre el contenido de cada una de las comunidades identificadas. No obstante, por sí solas no nos permite monitorear la disposición de causas a las que se busca hacer seguimiento o cuantificar la dimensión de las comunidades.

Dado los puntos anteriores, es que consideramos que la mejor forma de reportar los resultados de asociación para la búsqueda y monitoreo de causas asociadas es a través de lo que denominados *Tablas de reporte general*. En estas se integran las funcionalidades de las visualizaciones de grafos y Word Clouds, ya que como se aprecia en la Figura 5.13[2], es posible cuantificar (y de forma exacta) la cantidad de causas en un cluster, así como la cantidad de causas del grupo de estudio en el mismo conjunto. Finalmente permite visualizar cuales son los términos más relevantes desde un punto de vista de su aparición en los clusters.

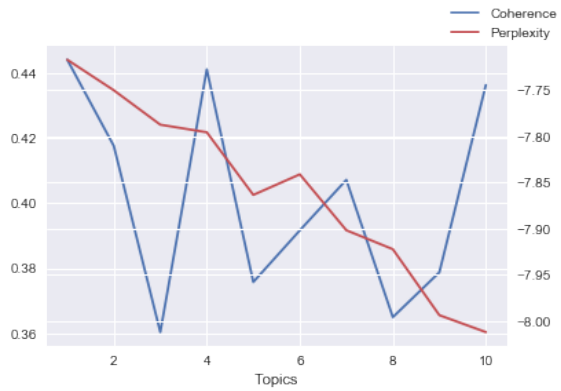
Visualización de tópicos

Finalmente, y cómo se aprecia en la Figura 5.13, hay veces en las que las comunidades o cluster que se identifican en la recuperación de documentos, siguen siendo muy extensas como para poder manejarlas o en este caso, investigar de manera conjunta. Es por esto que proponemos un tercer paso en este sistema de asociación de causas que es el de buscar tópicos dentro de los cluster más numerosos[3]. Para esto utilizaremos Latent Dirichlet Allocation (LDA) cómo modelo para buscar tópicos dentro de estas comunidades.

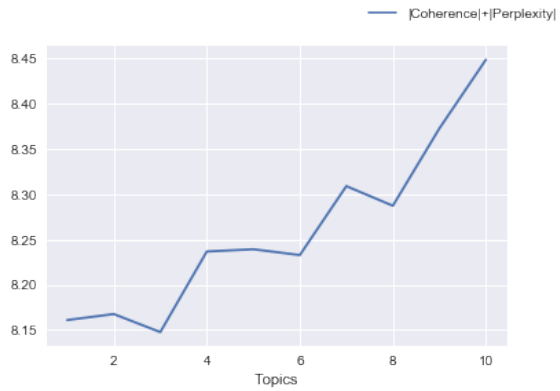
Si bien LDA es un modelo ampliamente utilizado y validado en la literatura, este requiere del seteo de un hiper parámetro, que en este caso es el número de tópicos que se buscarán en el corpus. Para esto existen varias métricas que se utilizan frecuentemente para analizar la 'calidad' de los tópicos encontrados: Topic Coherence y Perplexity. Estás medidas han sido

[2]Esta tabla da cuenta del resumen de las métricas que ahí se aprecian para una selección de causas basado en una similitud igual o superior a 0.8, según lo computado por Doc2Vec.

[3]Mencionamos con anterioridad que dada la extensión de media de las causas de este corpus, utilizar modelos de topicalización sólo tiene sentido si existen más de 50 causas en un corpus, ya que de nuestro conocimiento, Latent Dirichlet Allocation no ha sido testeado en corpus de menos de 10.000 palabras.



(a) Coherence y Perplexity para un rango de 10 tópicos



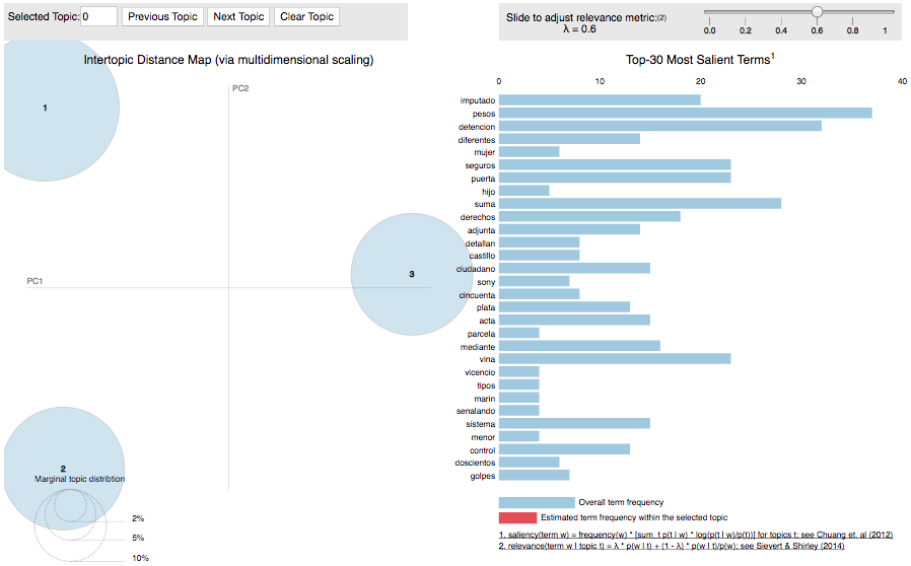
(b) Diferencias entre los valores de Coherence y Perplexity

Figura 5.14: Computo de Coherence y Perplexity para la selección del Nº óptimo de tópicos

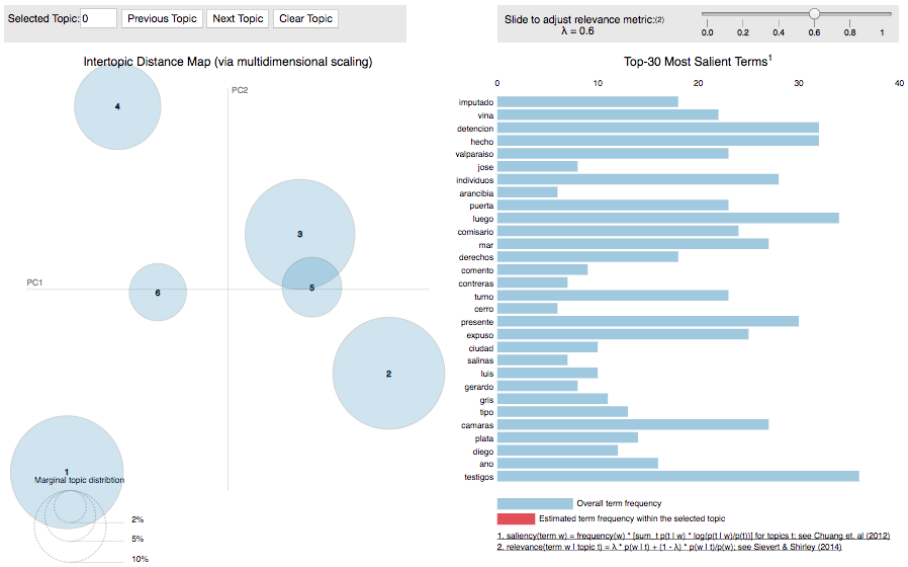
introducidas en el capítulo 2 y por ahora sólo nos restringimos a resumir que se busca el mínimo valor tanto de Topic Coherence (donde su dominio es negativo, por tanto, mientras más cercano a 0 mejor mejor es la calidad de tópico) como de Perplexity (cuyo valor es positivo, y por tanto también se busca el valor más cercano a 0). En la Figura 5.14 se puede apreciar al lado izquierdo el valor de ambas medidas en un rango de 1 a 10 tópicos para la muestra y al lado derecho de la Figura, la suma en términos de valor absoluto para estas dos métricas. Dado que buscamos que ambos valores sean lo más cercanos a 0, esto también aplica para la suma en término de valor absoluto para ambos, por tanto diremos que para este caso, 3 es el número óptimo de tópicos.

Por otro lado, y en línea con la búsqueda de entender de qué tratan los cluster o en este caso tópicos, aprovechamos la implementación de la librería pyLDAvis para visualizar las palabras con mayor probabilidad de ocurrencia dentro de los tópicos y la disposición espacial espacial de los tópicos para visualizar la diferencia o sobreposicionamiento entre tópicos en términos de las palabras que lo componen. En la Figura 5.15 se pueden apreciar cómo la visualización muestra los términos más probables de ocurrir y una disposición espacial de los tópicos. Basta observar la diferencia en la posición de los tópicos en términos espaciales para observar que con 3 tópicos existe menos superposición de términos que en el caso de 6 tópicos.

Con esto concluimos la sección de la visualización de los resultados concluyendo que es posible llegar un nivel de bastante granularidad para la asociación de causas, permitiendo facilitar la labor de comunicación sobre el contenido de los grupos conformados a través del proceso.



(a) Visualización para 3 tópicos



(b) Visualización para 6 tópicos

Figura 5.15: Visualización de tópicos con pyLDAvis

5.2. Seguimiento del Foco de Investigación

La segunda parte de los resultados de este trabajo es un seguimiento del ya descrito Foco Nº23 de la Fiscalía de San Antonio. Esto quiere decir que dada las causas que ya existen en el foco, se intentará buscar causas que sean similares y por tanto candidatos a ser incorporados en el foco. Para esto, dado que no se pretende constituir un foco a partir de un tema, sólo se seleccionarán las causas con mayor nivel de similitud textual a cada una de los casos ya incorporados en el foco y por tanto, se intentará validar que existe una correspondencia entre el nivel de similitud computado por Doc2Vec y lo que un experto considera como similitud para 2 delitos.

5.2.1. Selección de la muestra

Dado que analizar delitos es un proceso extenso, es que fue necesario seleccionar una muestra lo más acotada posible tal que nos permitirá validar nuestro supuesto de relevancia para vincular causas a través de similitud semántica textual, es por esto que se acordó seleccionar 7 de las causas que poseían imputado desconocido en el Foco de Investigación y para cada una de esas causas, seleccionar las 10 causas más similares en términos de similitud semántica textual. En la Tabla 5.3 se puede apreciar una representación de las causas que fueron entregadas a los analistas del Ministerio Público que han constituido el foco original. La Tabla muestra que para cada uno de las causas caracterizados por su Rol Único de Causa (RUC) existe una selección de 10 causas a las cuales se busca validar su similitud.

Para validar la similitud, y como fue descrito en los objetivos de la validación por juicio experto, se buscaba que cada analista analizará las causas de a pares $(RUC^{foco}, RUC^{foco}_{i\in(1,..,10)})$ y respondieron la pregunta: ¿Poseen estas 2 causas un *modus operandi* o caracterización del delincuente similar? Las respuesta válidas para esta pregunta sólo podrían ser SI o NO. Se escogió esta metodología por sobre una escala Likert (Likert, 1932) precisamente por los problemas que han sido ampliamente reportadas sobre esta escala. Algunos problemas reportados en la literatura son[4]: (1) Existe una fuerte tendencia hacia la respuesta central ya que es considerada una respuesta 'segura'. Esto se conoce como central tendency bias, (2) Se asume que la representación es equidistante, es decir, hay tanta distancia entre estar de acuerdo con una afirmación cómo a estar muy de acuerdo, lo cual también es un supuesto fuerte y cuestionable pues cada persona posee diferente estándares para escoger una respuesta, (3) No es posible generar estadísticos a través de las respuestas. Algo usual que se hace de forma incorrecta es cambiar la codificación de las respuestas y por ejemplo, asignar un 1 a Muy en desacuerdo. Esto no es válido, ya que la respuesta es ordinal y no poseen métricas de distancia.

Un punto interesante de las causas entregadas, es que si bien se buscaba analizar 70 comparaciones entre RUC, la cantidad única de RUC más similares a las causas de Foco

[4]Buenas fuentes bibliográficas para conocer más sobre la escala Likert y sus usos son: Cohen, L., Manion, L., y Morrison, K. (2000). Research methods in education (5th ed.). New York: Routledge. (pp. 253-255), Gilbert, G. N. (2008). Researching social life (3rd ed.) y Jamieson, S. (2004). Likert scales: how to (ab) use them. Medical education, 38(12), 1217-1218.

RUC Foco Investigación	$RUC_{i,1}$	...	$RUC_{i,10}$
RUC_1	$RUC_{1,1}$	...	$RUC_{1,10}$
RUC_2	$RUC_{2,1}$	...	$RUC_{2,10}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
RUC_7	$RUC_{7,1}$	...	$RUC_{7,10}$

Tabla 5.3: Representación de la base entregada a analistas de Ministerio Público

de Investigación eran 38. Es decir, habían causas en el Foco de Investigación que podían ser vinculadas a través de una tercera causa que no se encontraba en el foco, lo cual dice bastante sobre la eventual pertinencia de esas causas.

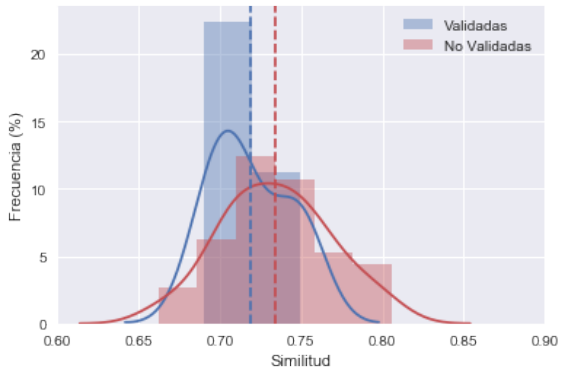
5.2.2. Resultados del seguimiento al Foco de Investigación

De las 70 comparaciones que fueron entregadas para analizar, se poseen 56 respuestas válidamente emitidas. A 9 de estas comparaciones la respuesta fue Sí a la pregunta: ¿Poseen estas 2 causas un *modus operandi* o caracterización del delincuente similar?. Por complemento, 47 de las respuestas fue No. Lo anterior nos dice que las recomendaciones basadas en Doc2Vec tuvieron **Precision = 19 %**. Respecto a la distribución de asociaciones validadas, es importante señalar que 4 de los 7 RUCs en estudio obtuvieron al menos 1 vínculo validado por los analistas (sumando entre esos 4 RUCs las 9 recomendaciones validadas).

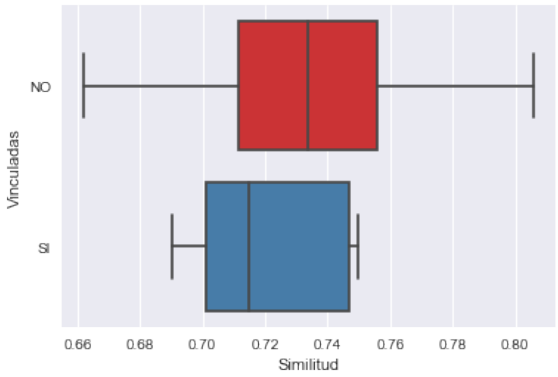
En lo siguiente, nos interesará analizar la correspondencia entre la similitud computada y la validación de los analistas, siendo los esperado que aquellas asociaciones que fueron validadas tengan niveles de similitud semántica textual superiores a las no validadas (o rechazadas). En la Figura 5.16 se puede apreciar las distribuciones para las similitudes computadas según la categoría validada o no validada. Un hecho contraproducente es que el promedio de las similitudes entre causas para el grupo de causas cuyo vínculo fue validado es de 0.71, mientras que el promedio para los vínculos no validados es de 0.73, lo cuál es totalmente contrario a los esperado. No obstante al testear mediante un test de medias, no es posible rechazar la hipótesis nula de que ambos promedios sean iguales ($p - value = 0{,}18$). Así mismo, al intentar constatar si ambas distribuciones son independientes mediante un test de Kolmogorov-Smirnov, tampoco fue posible rechazar la hipótesis de que ambas distribuciones son iguales ($p - value = 0{,}28$). Por tanto, si bien el resultado es contraintuitivo, no es verificable estadísticamente. No obstante, es necesario destacar que la cantidad de muestras para ambos grupos, los validados y los no validados es sumamente baja (9 y 47 registros respectivamente).

5.2.3. Visualización de los resultados

Algo interesante de visualizar es cómo las causas que fueron recomendadas por el algoritmo se vinculan de forma múltiple con los RUCs del Foco de Investigación, esto ya que sin supervisión alguna, las métricas de similitud indicaron que existían causas fuera del Foco de



(a) Histograma y KDE de la distribución de similitud entre causas

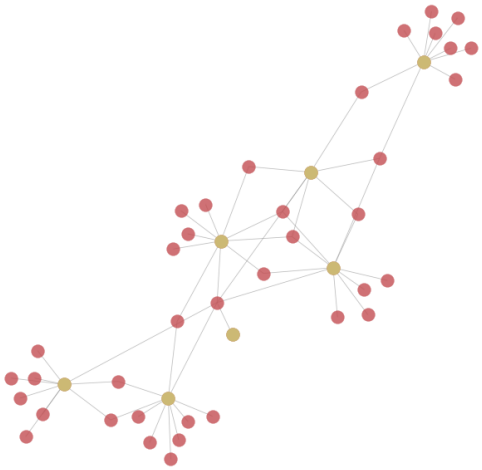


(b) Boxplot de la similitud entre causas

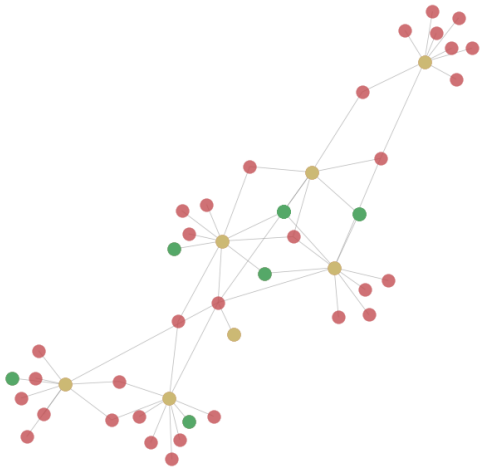
Figura 5.16: Distribución de la similitud entre causas del foco y las 10 más similares

Investigación que tenían alta similitud de forma simultánea con más de una causa del Foco de Investigación.

En la Figura 5.17 (c) se puede ver en verde aquellos RUCs cuya vinculación fue validada a alguna causa contenida en el Foco de Investigación (nodos en amarillo). Existe un nodo en verde que está conectado a 3 causas del Foco de Investigación, lo que inmediatamente nos permite tener una noción de la relevancia de esa causa.



(a) Causas del foco (amarillo) y similares (rojo)



(b) Causas del foco (amarillo), validadas (verde) y no validadas (rojo)

(c) Causas del foco (amarillo) y las validadas (verde)

Figura 5.17: Representación en grafo de las causas del foco y las más similares

Capítulo 6

Conslusiones

En la asociación de causas por similitud el proceso comenzó con un conjunto de 3.803 causas y el ejercicio fue analizar la factibilidad converger a través de la búsqueda por queries y similitud semántica textual al Foco de Investigación Nº23 que constaba con 16 registros (es decir, la idea era recuperar un conjunto que representa un 0,42 % de universo total sin más que una query y la optimización de 2 parámetros). Los resultados fueron $Recall : 4/16 = 25\,\%$, $Precision : 4/62 = 6{,}45\,\%$ y $F1 - Measure : 1/62 = 1{,}61\,\%$. Los resultados no lucen del todo bien, no obstante es necesario comprender que esto puede ser descrito como un problema en extremo desbalanceado y que tal como se comprobó en el seguimiento del Foco de Investigación, no cuenta con un *ground truth*, ya que causas que no estaban en el Foco, si podrían se parte de éste en términos semánticos. De hecho, una fuerte recomendación para trabajos futuros es la búsqueda de métricas que permitan evaluar de mejor forma los resultados de estas asociaciones. Finalmente, queremos destacar que incluso ante lo adverso del ejercicio, el modelo fue capaz de recuperar un 25 % de las causas, lo cual es el resultado de un proceso robusto con objetivos y tareas bien definidas, lo que permite pensar que efectivamente este parece ser un buen camino para abordar este problema.

En cuanto a los resultados del seguimiento del Foco de Investigación, podemos señalar que de las 56 respuestas válidamente recolectadas, 9 asociaciones fueron validadas y 47 descartadas, lo que indica que se obtuvo un $19,15\,\%$ en términos de *Precision*. Por otro lado y desafortunadamente, no es posible concluir respecto a si existe una correspondencia entre la similitud computada y la forma de comisión de un delito. En caso de obtener un resultado para esta métrica hubiese sido posible fijar un nivel de similitud semántica óptimo en términos de esos resultados para futuras recomendaciones. De todas formas, sí podemos señalar que ninguna asociación de causas fue validada con una similitud inferior a 0,69. Lo anterior podría servir como un input para el diseño de futuros seguimiento a Focos de Investigación. Finalmente, que se haya validado el 16 % de la muestra también nos indica que existen potencial para este método y que es capaz de encontrar asociaciones entre causas.

Respecto a los modelos implementados, fue posible constatar la rapidez de su ejecución y las radicales diferencias en los resultados de las métricas de similitud computados por LSI y Doc2Vec. Por otro lado, queremos destacar que incluso cuando los resultados en cuanto a métricas de desempeño no son sorprendentes, este sistema permite tener una priorización en

la investigación de causas, ya que para constituir un Foco de Investigación en la actualidad, las causas son muestreadas muchas veces de forma aleatorio. En cambio, nuestro sistema permite analizar 3.803 causas y obtener un conjunta de delitos agrupados y con información respecto a los tópicos que tratan las causas contenidas en sus conjuntos en menos de 3 minutos. Esto representa sin dudas progreso gigantesco en la capacidad de analizar causas por parte del Ministerio Público.

6.1. Recomendaciones para trabajos futuros

Este trabajo está lejos de ser una herramienta lista para ser implementada en el Ministerio Público, pero sí creemos que la perspectiva desde lo cual lo aborda es la correcta. Por tanto, dejaremos expresadas aquellas brechas que creemos es necesario acortar entre estre trabajo y una solución definitiva:

1. **Eliminar información irrelevante de los relatos**: gracias a los análisis de Patricio Moya, quién es integrante de la comisión de este trabajo, fue posible descubrir que las denuncias poseen una estructura muy similar, en donde primero se da cuenta de información respecto a cuándo y quién realiza la denuncia, luego se relatan los hechos que caracterizan al delincuente y su actuar, para finalmente describir la existencia de especies, testigos y cámaras, así como formalidades respecto al procedimiento que sucede a la denuncia. En términos de de asociación de delitos podría ser interesante la hora, lugar, día y caracterización de la víctima como tipología, sin embargo, lo que más contribuye para asociar delitos se encuentra en el relato del hecho. Por tanto se recomienda buscar una forma de extraer sólo la información relacionada con el suceso para su posterior análisis.
2. **Implementar modelos alternativos**: las condiciones en que se desarrolló este trabajo llevaron a cometer ciertas inconsistencias dinámicas, dentro de las cuales la carencia de modelos como *baseline* es uno de estas inconsistencias. Probablemente modelos más sencillos cómo TF-IDF o Latent Semantic Analysis, podrían por si sólos no entregar muy buenos resultados, pero es necesario constatar ese hecho, al menos para entender que tan bien lo hacen los modelos implementados u otros que se quieran analizar.
3. **Experimentar con distintas métricas de similitud**: en los dos modelos de similitud expuestos en este trabajo se utilizó la similitud coseno para cuantificar la similitud entre dos vectores que representan documentos. No obstante, existe la posibilidad de que otras medidas de distancia entreguen resultados interesantes de analizar.
4. **Construir una base de datos con más registros**: este punto es definitivamente muy relevante, y es que con la base de datos actual no se pueden concluir resultados realmente significativos. Ya fue posible observar como el seguimiento del foco estuvo exento de conclusiones respecto a la relevancia de la similitud semántica textual por lo acotado de sus observaciones, esto mismo se expresa en muchos otros aspectos del trabajo como el desempeño de los modelos o la ausencia de atributos que permitan sacar provecho a las características particular de diferentes delitos.
5. **Complementar similitud semántica textual**: Dado que la similitud semántica textual lleva los documentos a una matriz de distancias y eso puede ser a su vez llevado a

una matriz de adyacencia, es que existen un sin fin de otros modelos que podrían definitivamente mejorar los resultados. Por ejemplo, en los grafos del seguimiento del Foco de Investigación se puede apreciar que hay causas que están vinculadas con muchas otras, lo cual da indicios que de que es posible analizar aspectos como centralidad o existencia de comunidades en estas representaciones. Creemos que en este punto existe un gran potencial no sólo para mejoras, sino que también para investigación e innovación.

Finalmente es bueno destacar que en el Ministerio Público existe un inmenso espacio de oportunidades para desarrollar modelos cuantitativos que permitan apalancar la productividad de esta institución en términos de la calidad y cantidad de las investigaciones que realizan a través de técnicas como las expuestas en este trabajo. La Fiscalía al ser la institución encargada de dirigir las investigaciones, posee toda la información de las denuncias ejercidas por ciudadanos de este país, por lo que las probabilidades para encontrar patrones y vínculos que permitan enjuiciar a los responsables de los delitos aumentan en la medida en que se denuncien los delitos y se implementen sistemas que permitan el análisis automatizado de las grandes cantidades denuncias que llegan a esta institución. Cabe destacar que en ningún momento se ha propuesta el despido o reemplazo de analistas y fiscales y reemplazarlos por sistemas. Al contrario, creemos que el juicio humano seguirá siendo el predominante y más certero método de vinculación de delitos por los próximas años, pero la limitada productividad de los humanos para analizar delitos está generando un banco de causas y víctimas con lo que el sistema está en deuda.

Bibliografía

[1] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., y Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.

[2] Alvarez, Jon Ezeiza y B, Hannah. (2017). A review of word embedding and document similarity algorithms applied to academic text. University Of Freiburg.

[3] Beccaria, C. (1764). On crimes and punishment,(trans. H. Pallouci). Indianapolis: Bobbs-Merrill.

[4] Becker, G. S. (1968). Crime and punishment: An economic approach. In The economic dimensions of crime (pp. 13-68). Palgrave Macmillan, Londres.

[5] Bengio, Y., Ducharme, R., Vincent, P., y Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155.

[6] Bennell, C., y Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. Science and Justice, 42, 153–164.

[7] Bennell, C., y Jones, N. J. (2005). Between a ROC and a hard place: A method for linking serial burglaries by modus operandi. Journal of Investigative Psychology and Offender Profiling, 2, 23–41.

[8] Bennell, C., Snook, B., MacDonald, S., House, J. C., y Taylor, P. J. (2012). Computerized crime linkage systems: A critical review and research agenda. Criminal Justice and Behavior, 39(5), 620-634.

[9] Bennell, C., y Canter, D. V. (2017). Linking Commercial Burglaries by Modus Operandi: Tests Using Regression and ROC Analysis. Science and Justice.

[10] Bentham, J. (1843). Principles of penal law. W.Tait.

[11] Biderman, A. D., y Reiss Jr, A. J. (1967). On exploring the"dark figure.ºf crime. The Annals of the American Academy of Political and Social Science, 374(1), 1-15.

[12] Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[13] Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

[14] Borg, A., Boldt, M., Lavesson, N., Melander, U., y Boeva, V. (2014). Detecting serial residential burglaries using clustering. Expert Systems with Applications, 41(11), 5252-5266.

[15] Candolle, A. (1987). Considérations sur la statistique des délits. Déviance et société, 11(4), 352-355.

[16] Carrara, F. (1991). Programa de Derecho Criminal, Edit. Departamento de Publicaciones Facultad de Jurisprudencia, UNL, Loja.

[17] Chen, M. (2017). Efficient vector representation for documents through corruption. arXiv preprint arXiv:1707.02377.

[18] Cherven, K. (2013). Network graph analysis and visualization with Gephi. Packt Publishing Ltd.

[19] Chi, H., Lin, Z., Jin, H., Xu, B., y Qi, M. (2017). A decision support system for detecting serial crimes. Knowledge-Based Systems, 123, 88-101.

[20] Cid Moliné, J., y Larrauri Pijoan, E. (2001). Teorías criminológicas. Explicación y prevención de la delincuencia. Bosch, Barcelona.

[21] Clarke, R., y Eck, J. E. (2014). Become a problem-solving crime analyst. Willan.

[22] Clarke, R. V. G., y Webb, B. (1999). Hot products: Understanding, anticipating and reducing demand for stolen goods (Vol. 112). Londres: Home Office, Policing and Reducing Crime Unit, Research, Development and Statistics Directorate.

[23] Cohen, L. E., y Felson, M. (2016). Social Change and Crime Rate Trends: A Routine Activity Approach (1979). In Classics in Environmental Criminology (pp. 203-232). CRC Press.

[24] Crossley, S. A., Dascalu, M., y McNamarac, D. S. (2017). How Important Is Size? An Investigation of Corpus Size and Meaning in both Latent Semantic Analysis and Latent Dirichlet Allocation. In 30th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017. AAAI Press.

[25] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., y Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

[26] Douglas, J. E., Ressler, R. K., Burgess, A. W., y Hartman, C. R. (1986). Criminal profiling from crime scene analysis. Behavioral Sciences y the Law, 4(4), 401-421.

[27] Dumais, S. T. (2004). Latent semantic analysis. Annual review of information science and technology, 38(1), 188-230.

[28] Egger, S. A. (1984). A working definition of serial murder and the reduction of linkage blindness. Journal of police science and administration, 12(3), 348-357.

[29] Farrington, D. P., McGee, T. R., Moyle, W., Khanal, S., Saufi, A., Reid, S., ... y Lewandowski-Cox, N. (2017). The Integrated Cognitive Antisocial Potential (ICAP) theory: Empirical testing. Routledge International Handbook of Lifecourse Criminology.

[30] Felson, M. (1994). Crime and everyday life: Insight and implications for society. Thousand Oaks, CA: Pine.

[31] Garrido, V., Stangeland, P. y Redondo, S. (2006). Principios de Criminología, revisada y ampliada. Valencia: Tirant lo Blanch.

[32] Gil, D. B. (2016). ¿Qué es la criminología?: Una aproximación a su ontología, función y desarrollo. Derecho y Cambio Social, 13(44), 1.

[33] Gilbert, N. (2008). Researching Social Life 3rd Edition.

[34] Gwinn, S. L., Bruce, C. W., Hick, S. R., y Cooper, J. P. (Eds.). (2008). Exploring crime analysis: Readings on essential skills. International Association of Crime Analysts.

[35] Hagan, F. E. (2010). Introduction to criminology: Theories, methods, and criminal behavior. Sage.

[36] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[37] Huang, E. H., Socher, R., Manning, C. D., y Ng, A. Y. (2012, July). Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 873-882). Association for Computational Linguistics.

[38] Illescas, S. R. (2015). El origen de los delitos: introducción al estudio y explicación de la criminalidad. Tirant lo blanch.

[39] Jakobs, G. (2006). La pena estatal: significado y finalidad. Traducido y editado en Madrid.

[40] Jacomy, M., Venturini, T., Heymann, S., y Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software.

[41] Jelinek, F., Mercer, R. L., Bahl, L. R., y Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America, 62(S1), S63-S63.

[42] Le, Q., y Mikolov, T. (2014, January). Distributed representations of sentences and documents. In International Conference on Machine Learning (pp. 1188-1196).

[43] Likert, R. 1932. A technique for the measurement of attitudes. Archives of Psychology. 22(140).

[44] Lilly, J. R., Cullen, F. T., y Ball, R. A. (2010). Criminological theory: Context and consequences. Sage.

[45] Machicado, J. (2010). Concepto de delito. Apuntes Juridicos. Recuperado en Junio de 2018.

[46] McCulloch, W. S., y Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

[47] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[48] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., y McCallum, A. (2011, Julio). Optimizing semantic coherence in topic models. En Proceedings of the conference on empirical methods in natural language processing (pp. 262-272). Association for Computational Linguistics.

[49] Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris.

[50] Nemhauser, G. L., y Wolsey, L. A. (1988). Integer and combinatorial optimization. Interscience series in discrete mathematics and optimization. ed: John Wiley y Sons.

[51] Peillard, A. M. M., Correa, N. M., Chahuán, G. W., y Lacoa, J. F. (2013). La Reincidencia en el Sistema Penitenciario Chileno.

[52] Pennington, J., Socher, R., y Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[53] Perry, J. W., Kent, A., y Berry, M. M. (1955). Machine literature searching x. machine language; factors underlying its design and development. American documentation, 6(4), 242-254.

[54] Porter, M. D. (2016). A Statistical Approach to Crime Linkage. The American Statistician, 70(2), 152-165.

[55] Reich, B. J., y Porter, M. D. (2015). Partially supervised spatiotemporal clustering for burglary crime series identification. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(2), 465-480.

[56] Rehurek, R., y Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

[57] Robinson, P. H., y Darley, J. M. (2004). Does criminal law deter? A behavioural science investigation. oxford Journal of Legal studies, 24(2), 173-205.

[58] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

[59] Rubner, Y., Tomasi, C., y Guibas, L. J. (1998, January). A metric for distributions with applications to image databases. In Computer Vision, 1998. Sixth International Conference on (pp. 59-66). IEEE.

[60] Rumelhart, D. E., Hinton, G. E., y Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533.

[61] Shaw, C. R., y McKay, H. D. (1969). Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquency in Relation to Differential Characteristics of Local Communities in American Cities. In Classics in Environmental Criminology (pp. 103-140). CRC Press.

[62] Shwartz-Ziv, R., y Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.

[63] Sievert, C., y Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

[64] Skjong, R., y Wentworth, B. H. (2001, Enero). Expert judgment and risk perception. In The Eleventh International Offshore and Polar Engineering Conference. International Society of Offshore and Polar Engineers.

[65] Tonkin, M., Grant, T., y Bond, J. W. (2008). To link or not to link: A test of the case linkage principles using serial car theft data. Journal of Investigative Psychology and Offender Profiling, 5(1-2), 59-77.

[66] Tonkin, M., Woodhams, J., Bull, R., y Bond, J. W. (2012). Behavioural case linkage with solved and unsolved crimes. Forensic Science International, 222(1-3), 146-153.

[67] Utkin, L. V. (2006). A method for processing the unreliable expert judgments about parameters of probability distributions. European Journal of Operational Research, 175(1), 385-398.

[68] Uysal, A. K., y Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing y Management, 50(1), 104-112.

[69] Wilbur, W. J., y Sirotkin, K. (1992). The automatic identification of stop words. Journal of information science, 18(1), 45-55.

[70] Woodhams, J., Hollin, C. R., y Bull, R. (2007). The psychology of linking crimes: A review of the evidence. Legal and Criminological Psychology, 12(2), 233-249.

Anexos

A. Esquematización del proceso desarrollado en este documento

B. Estructura de las denuncias

En la Figura 6.2 se puede apreciar la estructura típico de un relato, en donde se identifican 3 partes. Las formalidades donde se especifica: hora de la denuncia, lugar, víctima, procedimiento que da origen a la denuncia, etc. En la segunda parte de describe el relato del hecho, es decir, la forma de comisión del delito y descripción del delincuente si es que se conoce. Finalmente, se señala si hay testigos o cámaras, se especifican las especies sustraídas y se señala cuales son los procedimientos que suceden a la denuncia.

C. Recuperación de documentos sólo con Doc2Vec

Una duda válida es preguntarse qué ocurre si por opción se hubiese escogido escoger cómo método de búsqueda única y exclusivamente los métodos de búsqueda de similitud semántica textual. En la Figura 6.3 se muestra cómo evoluciona el grafo con dicha opción. Se puede apreciar que la configuración del grafo

D. Similitud computada con Doc2Vec en causas del Foco de Investigación y el conjunto de 3.803 causas

En la Figura 6.4 se puede apreciar que a simple vista la distribución de las similtudes

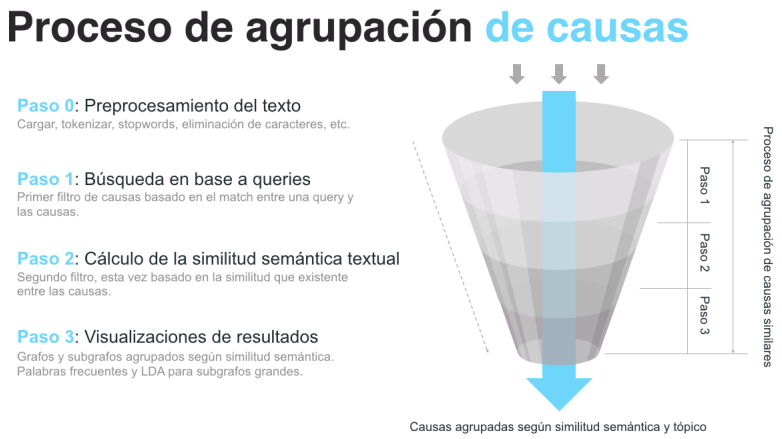


Figura 6.1: Secuencia de tareas para cumplir con el objetivo general

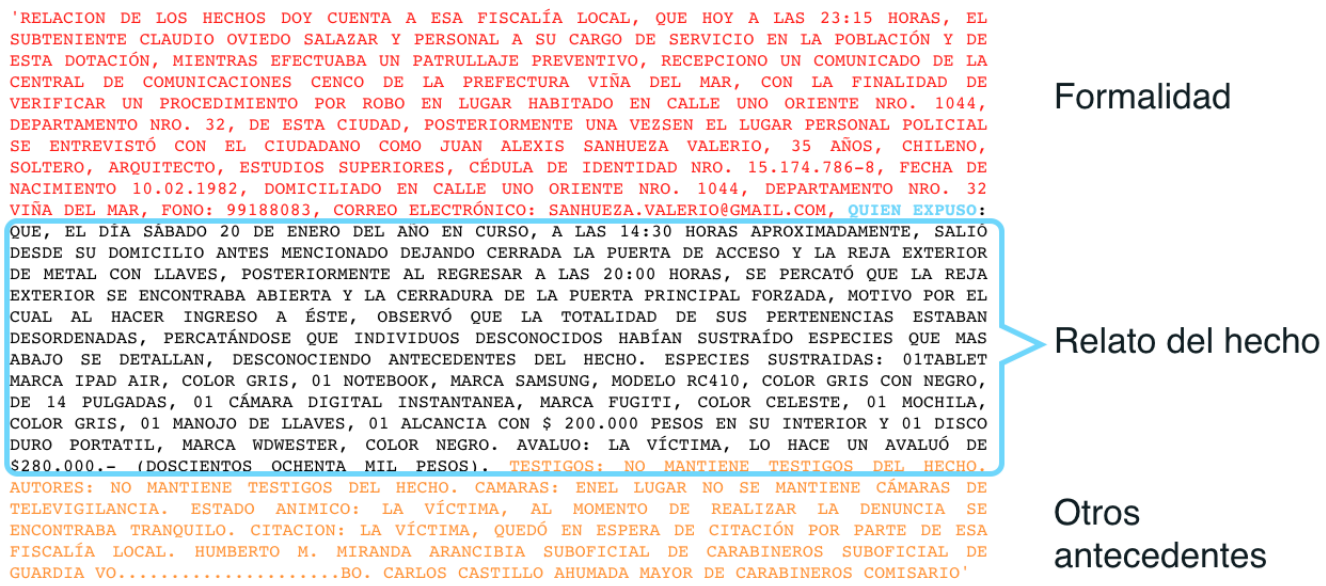
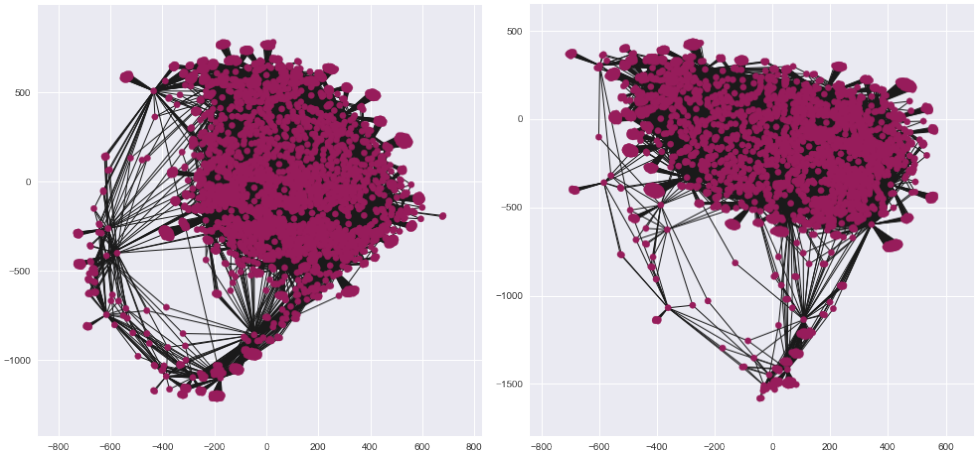


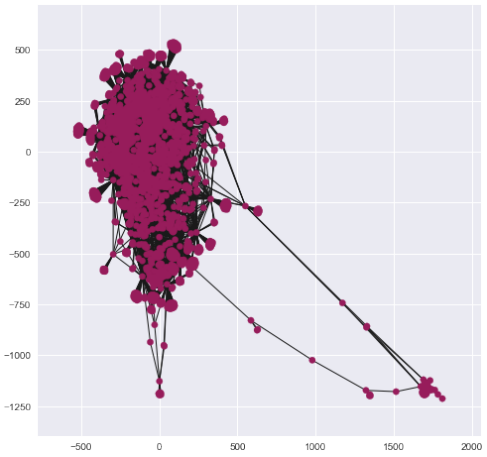
Figura 6.2: Proceso de persecución penal

entre todas las causas no difiere de la similitud computada entre las causas del Foco de Investigación, por lo que no podemos concluir que esta medida es por si sóla suficiente para encontrar causas que puedan constituir un Foco de Investigación.

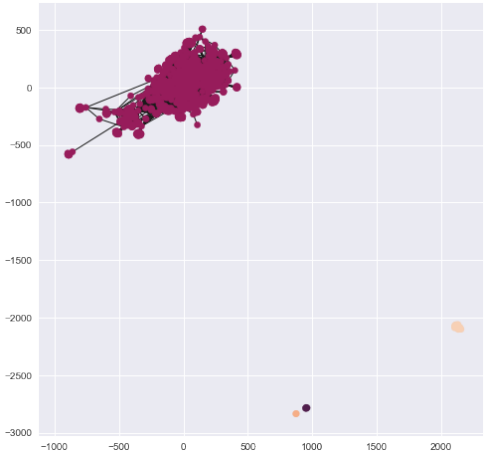


(a) Grafo para similitud = 0.4

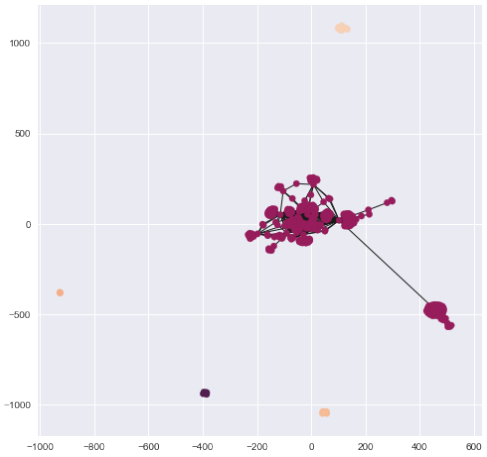
(b) Grafo para similitud = 0.5



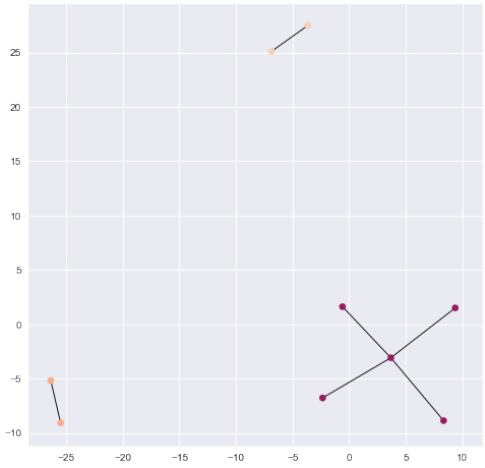
(c) Grafo para similitud = 0.6



(d) Grafo para similitud = 0.7



(e) Grafo para similitud = 0.8



(f) Grafo para similitud = 0.9

Figura 6.3: Cambios en la topología del grafo de documentos por nivel de similitud

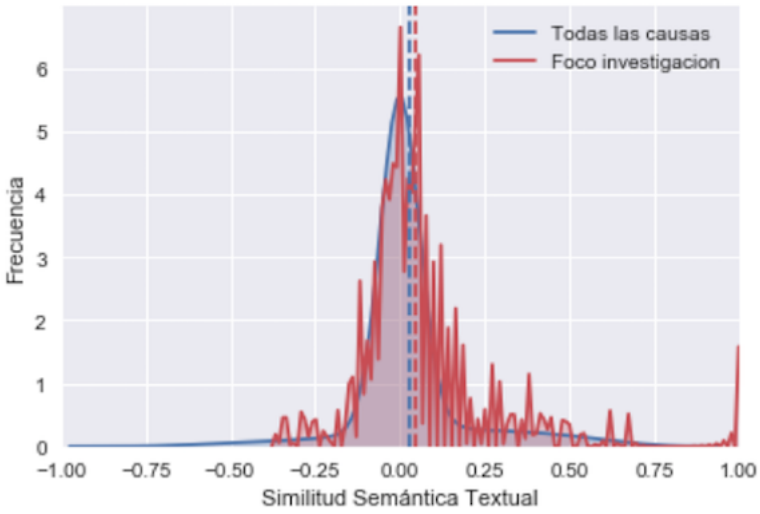


Figura 6.4: Similitud computada con Doc2Vec para todas las causas y el Foco de Investigación